

中图法分类号: TP751.1 文献标识码: A 文章编号: 1006-8961(XXXX)XX-0001-19

论文引用格式: Wang Yifei, Zhao Lian, Li Lei, You Shuhang, Liu Fan. Semantic dynamic boundary enhancement for referring remote sensing image segmentation: SDM-SAM2[J/OL]. Journal of Image and Graphics, XXXX: 1-19. DOI: 10.11834/jig.260290. (王一飞, 赵联, 李雷, 游书航, 刘帆. 语义动态边界增强的遥感图像指代分割方法: SDM-SAM2[J/OL]. 中国图象图形学报, XXXX: 1-19. DOI: 10.11834/jig.260290.) [DOI: 10.11834/jig.260290]

语义动态边界增强的遥感图像指代分割方法: SDM-SAM2

王一飞, 赵联, 李雷, 游书航, 刘帆*

太原理工大学人工智能学院, 太原 030024

摘要: 目的 遥感图像指代分割旨在根据自然语言指令生成目标区域的像素级掩码。然而, 通用分割基础模型迁移至遥感场景时, 仍面临深层跨模态交互过程中语义约束逐渐衰减、解码阶段层级查询信息利用不充分以及复杂目标边界监督不充分等问题。方法 针对上述问题, 提出一种用于遥感图像指代分割的语义动态边界增强方法—SDM-SAM2。该框架以SAM2(Segment Anything Model 2)为基础, 在图像编码阶段构建阶段感知文本记忆, 通过保存并检索不同阶段的文本信息, 以维持指代表达约束, 从而缓解深层语义传播中的语义衰减问题; 在Mask Decoder中引入任务自适应多层查询融合模块, 通过可学习层级权重对初始查询和不同解码层查询进行融合, 使浅层定位信息与深层语义表征共同参与掩码预测, 提升模型对复杂目标的定位与分割能力; 同时设计文本加权Sobel边界损失, 在连续前景概率图上提取可反向传播的边界响应, 并结合文本指定目标对目标轮廓施加显式约束。结果 在RefSegRS和RRSIS-D两个数据集上开展实验。完整模型在RefSegRS测试集上取得78.05%的平均交并比(Mean Intersection over Union, mIoU)和84.16%的总体交并比(Overall Intersection over Union, oIoU), 在RRSIS-D测试集上取得68.49%的mIoU和78.59%的oIoU。两个数据集上的消融结果表明, 阶段感知文本记忆、任务自适应多层查询融合和文本加权Sobel边界损失均能从不同环节改善基础模型; 其中, 阶段感知文本记忆在两个数据集上均取得最高的单模块mIoU。不同边界算子对比结果表明, Sobel算子取得了三种候选算子中最高的mIoU。结论 SDM-SAM2从语义保持、动态解码和边界建模三个方面增强了SAM2对遥感图像指代分割任务的适配能力, 能够改善复杂遥感场景下文本指定目标的定位和掩码预测效果。代码链接: <https://github.com/3038794185-cloud/SDM-SAM2>。

关键词: 遥感图像指代分割; SAM2 (Segment Anything Model 2); 视觉语言融合; 阶段文本记忆; 查询融合; 边界损失

Semantic dynamic boundary enhancement for referring remote sensing image segmentation: SDM-SAM2

Wang Yifei, Zhao Lian, Li Lei, You Shuhang, Liu Fan*

School of Artificial Intelligence, Taiyuan University of Technology, Taiyuan 030024, China

Abstract: Objective Referring remote sensing image segmentation (RRSIS) identifies and segments the ground object described by a natural-language expression. The target category and spatial extent are jointly determined by the image and text, requiring fine-grained alignment among language semantics, object appearance, spatial relations, and pixel boundar-

收稿日期: 2026-05-22; 修回日期: 2026-09-09

* 通信作者: 刘帆 liufan@tyut.edu.cn

基金项目: 国家自然科学基金(61703299); 山西省基础研究计划(202203021221094)

Supported by: National Natural Science Foundation of China(61703299); Shanxi Provincial Basic Research Program(202203021221094)

©中国图象图形学报版权所有

ies. Large scale variations, dense objects of the same category, complex backgrounds, and weak boundaries make this alignment difficult in remote sensing scenes. Although the Segment Anything Model 2 (SAM2) provides strong generic visual representations and mask decoding, its direct adaptation to RRSIS has three limitations: language guidance may weaken across hierarchical visual stages, the final decoder query may underuse complementary intermediate states, and region-overlap losses do not explicitly constrain boundaries. This study develops a semantic dynamic boundary enhancement framework, termed SDM-SAM2, to address these limitations in a unified architecture. **Method** The framework is built on a SAM2-based RRSIS baseline and contains three complementary components. The image and expression are jointly encoded by Bidirectional Encoder representation from Image Transformers 3 (BEiT-3), which produces a global classification-token representation, token-level text features, visual tokens, and a coarse mask-related representation. The Hiera backbone extracts hierarchical image features, while language information is introduced repeatedly into the visual hierarchy. The visual tokens and mask-related representation generate a coarse mask prompt, which the prompt encoder converts into dense and sparse embeddings. The global multimodal representation supplies sentence-level semantic guidance, whereas the token-level text sequence preserves word and relational information. The mask decoder receives the final image embedding, high-resolution image features, prompt embeddings, and text features, and progressively updates its output-token queries through text-image interaction. Thus, the global classification token is a semantic prompt, the token-level text features are language context, and the decoder queries are learnable states updated to predict the mask; these tensors cooperate but are not interchangeable. The first component is stage-aware text memory. At each Hiera stage, it retains the full token-level text state and its mask-aware pooled summary. The pooled current text, historical text summaries, and visual context determine the relevance of each historical state, and softmax produces normalized memory weights. These weights are applied to the complete historical text sequences, preserving word-level structure. A learned gate, estimated from the pooled current state, retrieved memory, and visual context, then combines the retrieved and current-text sequences. Its scalar value is broadcast over token and channel dimensions, enabling stage-dependent retention of historical language information. The second component is task-adaptive multi-layer query fusion. The initial output-token state and intermediate query states from successive decoder layers form a query set. A lightweight network maps each state to a scalar score, and normalized scores weight the initial and intermediate states to produce the final query. The weights are learned globally with the segmentation objective and represent task-level contributions of decoder depths, rather than sample-specific attention values. Thus, early prompt and localization information can complement the semantic and contour cues in deeper queries. The third component is a text-weighted Sobel boundary loss. Horizontal and vertical Sobel operators extract differentiable gradient-magnitude maps from the predicted probability and ground-truth mask. The sentence-level representation is projected to a sample-level boundary weight and spatially repeated to the mask resolution. This weight modulates the mean squared error between the two gradient maps, relating boundary supervision to the referring expression while constraining boundary discontinuity and contour displacement. The total objective combines cross-entropy, Dice, and boundary losses. Accordingly, text memory maintains encoder guidance, query fusion integrates decoder states, and Sobel supervision constrains the final boundary. Experiments use RefSegRS and RRSIS-D, two public datasets with different sample sizes, resolutions, category compositions, and expression distributions. RefSegRS contains 4,420 image-expression-mask triplets covering 14 categories at 512×512 pixels, whereas RRSIS-D contains 17,402 triplets covering 20 categories at 800×800 pixels. Metrics include mean intersection over union (mIoU), overall intersection over union (oIoU), and the proportion of samples whose intersection over union exceeds threshold K (Pr@K). Ablations evaluate all individual and combined components, while an operator comparison examines Sobel, Laplacian, and Prewitt boundaries under the same setting. **Result** On the RefSegRS test set, SDM-SAM2 obtains an mIoU of 78.05% and an oIoU of 84.16%. Compared with the SAM2-based RS2-SAM2 baseline, it improves mIoU by 4.15 percentage points and oIoU by 3.29 percentage points. The improvements at the stricter Pr@0.7, Pr@0.8, and Pr@0.9 thresholds are 6.05, 9.08, and 15.85 percentage points, respectively, indicating that the proposed method increases not only average overlap but also the number of high-quality predictions. On the RRSIS-D test set, the proposed model reaches an mIoU of 68.49% and an oIoU of 78.59%. Relative to RS2-SAM2, mIoU increases by 1.77 percentage points, while Pr@0.5, Pr@0.6, and Pr@0.7 increase by 2.39, 2.26, and 2.10 percentage points, respectively. The oIoU and Pr@0.9 values are slightly lower

than those of the baseline, showing that the benefit is not uniform across every metric but is concentrated in mean region quality and the low-to-high intersection-over-union threshold range. The ablations confirm different component roles. Among single-module variants, stage-aware text memory gives the highest mIoU on both datasets: 67.89% on RRSIS-D and 78.47% on RefSegRS. The complete configuration is most balanced on RRSIS-D, whereas RefSegRS combinations are not strictly additive, indicating dataset-dependent marginal and joint effects. With the same boundary-loss setting, Sobel reaches 67.72% mIoU on RRSIS-D, exceeding Laplacian at 66.83% and Prewitt at 67.49%. Qualitative results show fewer incomplete regions, disconnected boundaries, and spatial deviations. The full model has 961.05 million parameters, 26.24 million or approximately 2.8% more than the 934.81-million-parameter baseline. **Conclusion** SDM-SAM2 maintains stage-dependent language guidance, exploits complementary decoder queries, and introduces expression-conditioned boundary supervision. Results on both datasets show consistent mIoU gains and improvements in most threshold-based measures with a limited parameter increase. Cross-dataset ablations further reveal that module combinations can have different marginal effects under different data distributions. Overall, the framework improves semantic localization and boundary delineation while retaining the general SAM2 mask-decoding structure. The codes are available at: <https://github.com/3038794185-cloud/SDM-SAM2>.

Key words: referring remote sensing image segmentation; Segment Anything Model 2 (SAM2); vision-language fusion; stage-aware text memory; query fusion; boundary loss

论文引用格式: Wang Y F, Zhao L, Li L, You S H, Liu F. 2026. Semantic dynamic boundary enhancement for referring remote sensing image segmentation: SDM-SAM2. *Journal of Image and Graphics*, (): (王一飞, 赵联, 李雷, 游书航, 刘帆. 2026. 语义动态边界增强的遥感图像指代分割方法: SDM-SAM2. *中国图象图形学报*, ():) [DOI: 10.11834/jig.260290]

0 引言

遥感图像分割是遥感智能解译中的基础任务, 常用于城市管理、灾害监测、土地利用分析和环境监测。传统语义分割通常基于闭集类别假设, 也就是说, 模型只能预测训练阶段预先定义类别。真实应用并不总是这样清晰: 用户可能希望通过一句指代表达定位目标, 例如“河流旁边的农田”或“左下角靠近道路的建筑”。此时, 模型不仅要识别图像内容, 还需要理解自然语言指令中的目标类别、空间关系和上下文线索。因此, 遥感图像指代分割成为一个重要研究方向, 其目标是根据给定指代表达, 在遥感图像中输出对应目标的像素级掩码(Yuan 等, 2024)。

与自然图像指代分割相比, 遥感图像指代分割的困难主要体现在文本语义判别、目标定位和边界恢复三个相互关联的环节。首先, 遥感图像多为俯

视视角, 同类目标在外观、颜色和纹理上往往较为相似。当场景中同时存在多个候选目标时, 模型需要结合指代表达中的目标属性、空间位置和对对象关系, 建立文本语义与视觉区域之间的细粒度对应, 才能准确筛选目标(王丽安等, 2023)。其次, 遥感目标具有显著的尺度变化和旋转任意性, 部分目标空间占比较小且分布密集, 容易受到相邻实例和复杂背景的干扰。已有研究表明, 直接将自然图像指代分割方法迁移至遥感场景时, 模型对小目标和分散目标的处理能力仍然有限(Yuan 等, 2024); RMSIN (Rotated Multi-Scale Interaction Network) 进一步将多尺度表征和旋转建模视为遥感图像指代分割的关键问题(Liu 等, 2024a)。此外, 道路、建筑和船只等目标还可能具有细长、低对比度或碎片化轮廓, 其边界容易受到阴影、背景纹理和邻近目标的影响。上述特点使遥感图像指代分割不仅需要理解文本描述, 还需要将文本语义进一步转化为准确的目标定位和精细的轮廓恢复。

近年来, 基础分割模型为解决上述问题提供了新的技术路径。SAM (Segment Anything Model) 通过可提示分割任务和大规模掩码数据展示了较强的零样本分割能力(Kirillov 等, 2023), SAM2 (Segment Anything Model 2) 进一步将该能力扩展到图像和视频, 并利用流式记忆建模目标变化(Ravi 等, 2025)。然而, SAM 和 SAM2 主要接收点、框或掩码等视觉提示, 并未直接面向自然语言驱动的遥感目标分割。

当提示由视觉输入转变为指代表达时,模型还需要解析文本中的目标属性、空间位置和对象关系,并将其转换为可靠的视觉区域响应。EVF-SAM (Early Vision-Language Fusion for Text-Prompted Segment Anything Model) 表明,较早阶段的视觉—语言融合有助于提高文本提示分割能力 (Zhang 等, 2024); RS2-SAM2 (SAM2 for Referring Remote Sensing Image Segmentation) 则通过联合编码器、双向层级融合和伪掩码提示,将 SAM2 进一步适配至遥感图像指代分割任务 (Rong 等, 2026)。这些研究说明,基础模型适配遥感指代场景的关键不仅在于引入文本信息,还在于如何使语言约束持续服务于目标判别、区域定位和掩码生成。

尽管上述方法已经推进了 SAM 系列模型在遥感图像指代分割中的应用,但前述任务困难在 SAM2 框架中仍表现为三个相互关联的结构性问题。面对相似目标干扰时,模型需要持续保留文本中的属性、位置和关系信息;然而,文本特征在深层视觉—语言交互中不断吸收图像上下文,细粒度指代线索可能逐渐弱化。面对小目标和密集目标时,准确分割既依赖浅层查询中的原始提示与局部定位信息,也依赖深层查询形成的语义响应;仅使用末层查询进行预测,难以充分利用不同层查询状态的互补信息。面对细长、低对比度或碎片化边界时,交叉熵 (cross-entropy, CE) 损失和 Dice 损失主要优化预测区域与真实掩码 (ground truth, GT) 的整体重叠,对文本指定目标的局部轮廓缺少显式约束。上述问题分别反映出文本约束在语义理解、目标定位和边界恢复三个环节中的利用仍不充分。

针对上述问题,本文提出 SDM-SAM2 语义动态边界增强的遥感图像指代分割框架。该框架以 RS2-SAM2 为基础、以 SAM2 为分割基座,围绕文本约束在语义理解、目标定位和边界恢复过程中的持续作用,构建了编码、解码与监督相衔接的增强链路。这三个关键组成部分分别承担语义保持、信息聚合和边界优化功能,共同完成文本约束下由粗到精的逐级处理。

主要贡献如下:

1) 提出 SDM-SAM2 遥感图像指代分割框架,将遥感场景中的语义判别、目标定位和边界恢复需求,转化为编码阶段语义保持、解码阶段查询聚合和监督阶段边界约束组成的连续增强链路。

2) 针对编码阶段的语义保持,提出阶段感知文本记忆模块。该模块在视觉骨干网络中保存并检索不同阶段的文本状态,以维持目标属性、空间位置和对象关系等原始指代信息,有效缓解了深层跨模态交互中的语义衰减问题。

3) 针对解码阶段的查询聚合,提出任务自适应多层查询融合模块。该模块在 SAM2 的 Mask Decoder 中动态聚合初始提示查询与各解码层查询,使浅层局部定位线索与深层语义响应共同参与最终掩码预测,提升了模型对密集及多尺度目标的定位能力。

4) 针对监督阶段的边界约束,设计基于 Sobel 算子的文本加权边界损失。该损失将文本条件引入可微边界感知,通过掩码概率图提取可反向传播的边界响应,对特定目标的轮廓施加显式的几何约束,显著改善了遥感目标细长、碎片化边界的分割质量。

1 相关工作

1.1 语义分割与边界建模

语义分割要求模型为图像中的每个像素分配类别。FCN (Fully Convolutional Networks) 将分类网络改造成全卷积网络,使模型可以端到端地输出像素级预测 (Long 等, 2015)。U-Net 使用编码器-解码器结构,通过收缩路径获取上下文,再通过扩张路径恢复空间位置 (Ronneberger 等, 2015)。DeepLabv3+ 结合空洞卷积和轻量解码器,在多尺度上下文和边界恢复之间取得平衡 (Chen 等, 2018)。这些方法为遥感图像分割提供了基本网络范式,但它们大多仍建立在闭集类别假设之上。

在自然图像指代分割中,CRIS (CLIP-Driven Referring Image Segmentation) 利用 CLIP (Contrastive Language-Image Pre-training) 的视觉—语言知识推进文本到像素的对齐 (Wang 等, 2022); LAVT (Language-Aware Vision Transformer) 将语言信息注入视觉 Transformer 的中间层,以加强早期跨模态融合 (Yang 等, 2022); VLT (Vision-Language Transformer and Query Generation) 通过语言相关查询生成来建模不同理解视角 (Ding 等, 2021); ReSTR 则使用纯 Transformer 结构完成跨模态特征融合 (Kim 等, 2022)。这些工作为 RRSIS (Referring Remote Sensing Image Segmentation) 提供了重要参考,但遥感图

像中的俯视视角、小目标密集和尺度变化仍会带来额外困难。

遥感图像指代分割还需要进一步建立自然语言指令与图像区域之间的对应关系。与此同时,传统区域级监督信号对小目标、细长目标和复杂拓扑边界的约束仍然有限。Boundary Loss通过边界距离补充区域监督,适合类别高度不平衡的场景(Kervadec等,2019)。Gated-SCNN(Gated Shape CNNs)将形状信息作为单独分支,并用门控机制帮助模型关注边界相关表征(Takikawa等,2019)。本文的Sobel文本加权边界损失受这类边界建模思想启发,但其监督对象进一步指向由指代表达限定的目标轮廓。高分辨率遥感语义分割研究也指出,复杂背景和多尺度地物会造成分割边界模糊、区域破碎等问题,低层细节对高层语义特征具有边界连续性补充作用(李林娟等,2024)。

1.2 遥感图像指代分割

RRSIS提出RefSegRS数据集,并使用语言引导的跨尺度增强模块来改善多尺度视觉特征(Yuan等,2024)。RMSIN进一步关注遥感目标的尺度变化和旋转任意性,设计了多尺度交互和自适应旋转卷积,并构建了规模更大的RRSIS-D数据集(Liu等,2024a)。这些工作说明,遥感图像指代分割不仅需要文本语义理解,还需要处理遥感图像中的多尺度、任意方向和密集干扰等问题。

后续研究更重视细粒度视觉—语言对齐。FIANet(Fine-Grained Image-Text Alignment Network)将原始指代表达拆分为目标文本和位置文本,通过更细的跨模态对齐增强多模态表示(Lei等,2025)。CADFormer(Fine-Grained Cross-modal Alignment and Decoding Transformer)指出,粗粒度单向对齐会削弱文本目标与视觉区域之间的对应关系,并通过双向语义引导对齐和文本增强解码网络加强跨模态关系(Liu等,2025)。CroBIM-U从不确定性角度处理不同区域的跨模态对齐可靠性,并把更多语言约束放到高不确定区域(Sun等,2026)。这些研究共同表明,RRSIS的关键不是简单拼接图像和文本,而是让语言约束在不同尺度、不同区域和不同解码阶段持续发挥作用。

近期研究还从不同角度继续扩展RRSIS。SBANet(Scale-wise Bidirectional Alignment Network)强调尺度级双向对齐,用视觉到语言和语言到视觉

的信息流改善跨模态表示(Li等,2025);Multi-Scale Spatially-Guided Joint Prediction引入多尺度空间引导和目标-背景双流解码,以提升细粒度区分能力(Zhang等,2025);CSINet(Cross-View Semantics Interaction Network)通过远近视角协同建模来处理遥感目标的剧烈尺度变化(Yang等,2025a)。此外,NWPU-Refer提供了更大规模的RRSIS数据集和基准,说明该领域仍需要更强的泛化评估(Yang等,2025b)。

1.3 基础模型与SAM系列方法

SAM提出了可提示分割任务,并用大规模掩码数据训练出具有较强迁移能力的分割模型(Kirillov等,2023)。SAM2将这一思路扩展到图像和视频,并使用流式记忆结构支持视频目标分割(Ravi等,2025)。这些模型的优势在于分割先验强、泛化范围广。不过,自然语言理解和遥感场景适配并不是它们的原始设计重点。

从视觉—语言基础模型看,CLIP通过大规模图文对比学习建立开放词表视觉—语言对齐能力(Radford等,2021);BEiT-3(Bidirectional Encoder representation from Image Transformers 3)将图像、文本和图文对统一到多模态预训练框架中(Wang等,2023);Grounding DINO则展示了文本 grounding 对开放集目标定位的价值(Liu等,2024b)。这些工作共同说明,自然语言指令可以作为开放场景视觉理解的重要条件。多模态语义分割综述进一步将特征增强、特征融合和多层特征交互概括为提升多模态分割性能的重要方向(赵什陆和张强,2023),为层级文本保持和查询融合设计提供了相关研究背景。

已有工作开始把基础模型用于指代分割及遥感图像指代分割。EVF-SAM研究了文本提示下的SAM适配,并强调早期视觉—语言融合的重要性(Zhang等,2024)。RSRefSeg使用CLIP进行视觉—语言编码,再用SAM细化掩码(Chen等,2025)。RSRefSeg 2将任务拆成粗定位和细分割两个阶段,把CLIP的跨模态对齐能力和SAM的分割能力结合起来(Chen等,2026)。DiffRIS使用预训练文本到图像扩散模型的语义能力,增强RRSIS中的跨模态对齐(Dong等,2025)。RS2-SAM2进一步直接适配SAM2,使用联合编码器、双向层级融合和伪掩码提示完成遥感图像指代分割(Rong等,2026)。本文在这一基础上进一步关注深层语义保持、解码查询利

用和文本条件边界学习。

2 研究方法

提出SDM-SAM2遥感图像指代分割框架。给定遥感图像 I 和自然语言指代表达 T ,模型需要输出文本指定目标的像素级掩码 M 。该框架以RS2-SAM2为基础、以SAM2为分割基座,围绕文本约束由语义理解向目标定位和边界恢复的传递过程,构建编码、解码与监督相衔接的增强链路。

在编码阶段,阶段感知文本记忆模块保存并检索不同阶段的文本状态,维持目标属性、空间位置 and 对象关系等指代信息;在解码阶段,任务自适应多层查询融合模块聚合初始查询和不同解码层查询,使局部定位线索与深层语义响应共同参与预测;在监

督阶段,文本加权Sobel边界损失对文本指定目标的轮廓施加显式约束。三部分分别承担语义保持、信息聚合和边界优化功能,共同构成编码、解码与监督相衔接的处理链路。

2.1 整体框架

SDM-SAM2延续RS2-SAM2的整体输入形式,SDM-SAM2的整体框架如图1所示。模型同时接收SAM2图像输入、EVF图像输入、文本token以及文本attention mask。BEiT-3多模态编码器首先联合编码EVF图像与指代表达,得到视觉词元特征visual_feat、文本词元特征text_feat、分类词元(classification token, CLS)全局多模态特征feat和用于生成粗掩码提示的掩码词元mask token(Wang等,2023)。随后,mask token与visual_feat进行交互,并通过mask_product生成伪掩码提示。

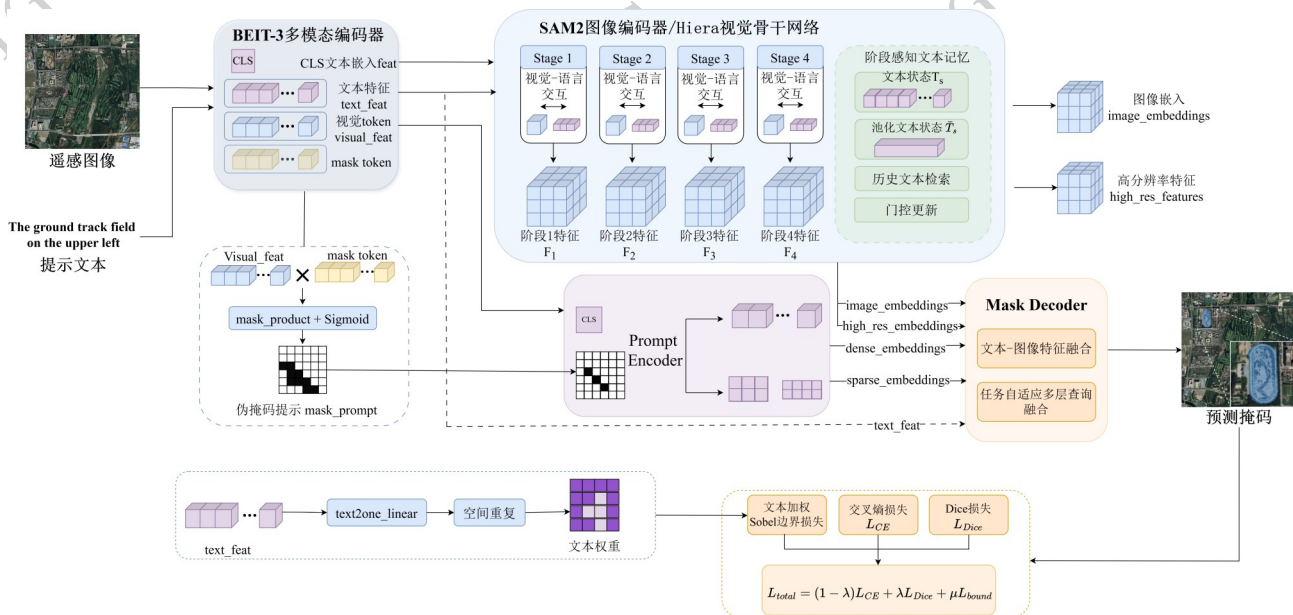


图1 所提出的SDM-SAM2框架的概览

Fig. 1 Overview of the proposed SDM-SAM2 framework

SAM2 图像编码器以遥感图像和 text_feat 为输入,在 Hiera 视觉骨干网络中进行多阶段视觉—语言交互(Ryali等,2023)。本文在该过程中插入阶段感知文本记忆模块,使文本状态不仅随当前视觉特征更新,也可以从历史阶段中检索更稳定的语义参考。Prompt Encoder接收CLS全局多模态特征 feat 和伪掩码提示,输出稀疏提示嵌入 sparse embeddings 与稠密提示嵌入 dense embeddings。Mask Decoder 进一步将图像嵌入 image_embeddings、高分辨率图像

特征 high_res_features 和提示嵌入结合,生成最终的掩码概率图。

2.2 阶段感知文本记忆模块

遥感图像指代分割中的文本指令通常同时包含目标类别、位置关系和上下文线索。例如“道路旁边的建筑”不仅要求模型识别建筑,还要求模型保留“道路旁边”这一空间约束。深层视觉—语言交互可以让文本特征吸收图像上下文,但也可能逐步稀释原始指代表达中的细粒度语义。为此,本文在视觉

骨干网络中引入阶段感知文本记忆模块,该模块的结构如图2所示。

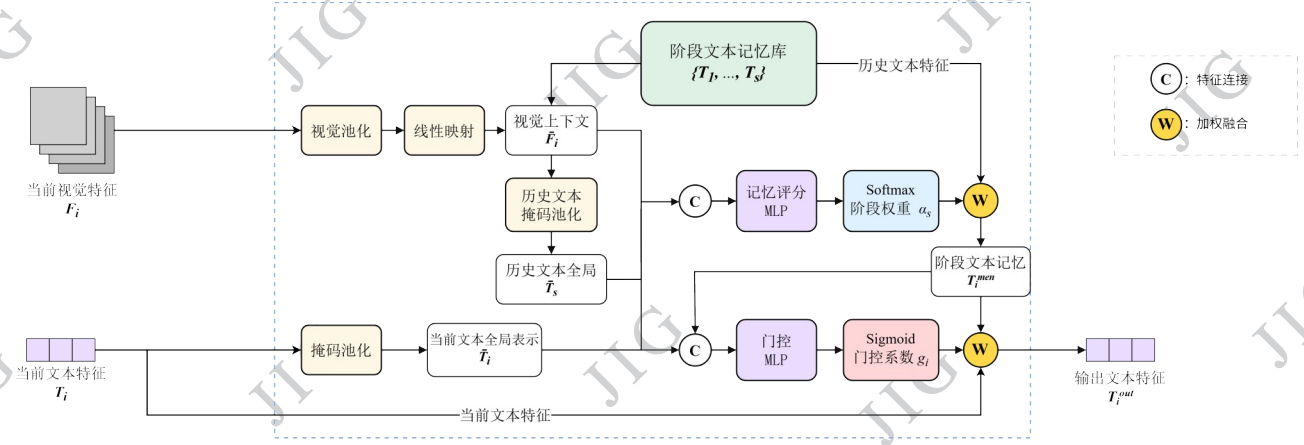


图2 阶段感知文本记忆模块

Fig. 2 Stage-Aware Text Memory Module

设第 i 个阶段的视觉特征为 F_i , 文本特征为 T_i 。模型在每个阶段末保存当前文本状态及其掩码平均池化表示, 形成历史阶段文本记忆集合。对于任意文本状态 T , 其有效 token 掩码为 m_n , 则池化表示写为:

$$\bar{T} = \frac{\sum_{n=1}^{N_i} m_n T_n}{\sum_{n=1}^{N_i} m_n + \varepsilon} \#(1)$$

式(1)中, $\bar{T}$ 表示掩码平均池化后的文本表示, N_i 表示文本 token 数量, T_n 表示第 n 个 token 的文本特征, m_n 表示第 n 个 token 的有效掩码值, ε 表示防止分母为零的极小常数。在当前阶段, 模块首先得到当前文本池化表示 $\bar{T}_i$, 同时对当前视觉特征 F_i 进行全局平均池化和线性投影, 得到视觉上下文表示 $\bar{F}_i$ 。随后, 模块对每个历史阶段记忆计算重要性分数:

$$a_s = MLP([\bar{T}_i, \bar{T}_s, \bar{F}_i]) \#(2)$$

$$\alpha_s = \frac{\exp(a_s)}{\sum \exp(a_j)} \#(3)$$

$$T_i^{\text{mem}} = \sum_s \alpha_s \bar{T}_s \#(4)$$

式(2)一式(4)中, a_s 表示第 s 个历史阶段记忆的重要性分数, $\bar{T}_s$ 表示第 s 个历史阶段的文本池化表示, α_s 表示归一化后的历史记忆权重, T_i^{mem} 表示当前阶段检索得到的历史文本记忆。直接用历史记忆替换当前文本状态可能带来过强干预。因此, 本文使用门控机制控制历史记忆的注入强度:

$$g_i = \sigma(MLP([\bar{T}_i, T_i^{\text{mem}}, \bar{F}_i])) \#(5)$$

$$T_i^{\text{out}} = (1 - g_i) \bar{T}_i + g_i T_i^{\text{mem}} \#(6)$$

式(5)和式(6)中, g_i 表示当前阶段的门控系数, $\sigma(\cdot)$ 表示 Sigmoid 函数, T_i^{mem} 表示门控残差更新后的输出文本特征。若当前阶段尚无历史记忆, 模块直接保留当前文本状态。当历史记忆可用时, 门控残差更新允许模型有选择地引入早期或中层文本语义, 从而缓解深层跨模态交互中的语义衰减。

2.3 任务自适应多层查询融合模块

SAM 系列方法在 Mask Decoder 中使用 Two-Way Transformer 交互更新提示查询和图像特征, 并通过最终注意力层生成用于掩码预测的查询表示 (Kirillov 等, 2023; Ravi 等, 2025)。在这一过程中, 查询向量是连接提示信息、视觉区域和最终掩码输出的关键载体。Transformer 的注意力机制本身以 query、key 和 value 为核心 (Vaswani 等, 2017), DETR (DEtection TRansformer) 也证明可学习查询向量可以作为目标级表示参与预测 (Carion 等, 2020)。因此, 如何更充分地利用查询向量, 会直接影响模型对遥感目标区域的定位与区分能力。原始 Two-Way Transformer 通常只将最后一层输出的查询向量送入最终注意力层。这个设计很简洁, 但浅层查询中保留的位置线索和初始提示信息, 可能在逐层更新后被削弱; 深层查询虽然具有更充分的语义交互, 却未必完整保留局部空间约束。多尺度视觉研究已经表明, 不同层级特征具有互补价值 (Lin 等, 2017; Zhu 等, 2021; Cheng 等, 2022)。受这些工作启发, 本文在 SAM2 的 Two-Way Transformer 中引入多层查询融合模块, 该模块的整体结构如图3所示。

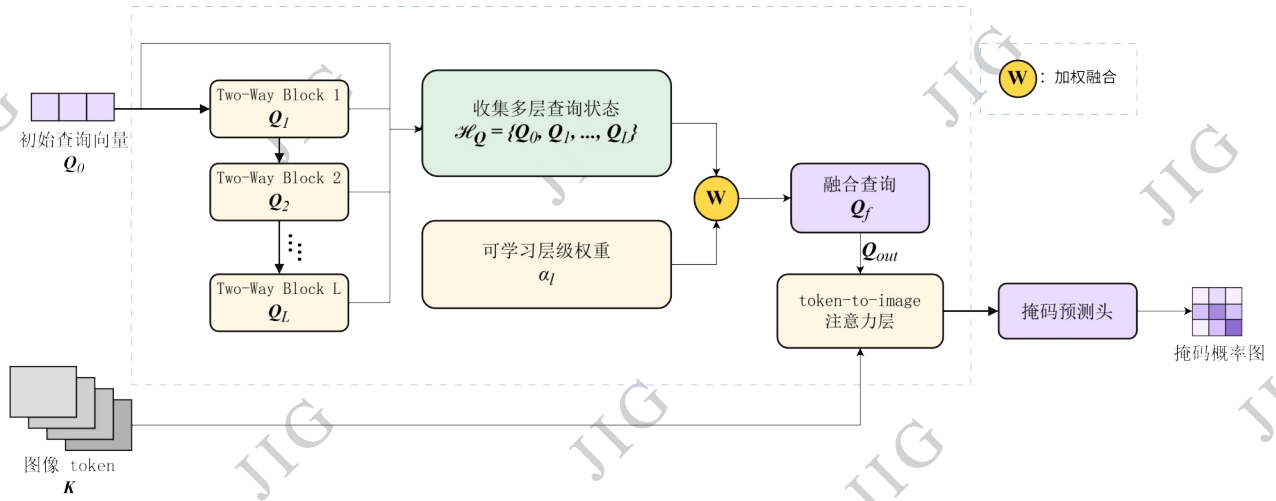


图3 任务自适应多层查询融合模块

Fig. 3 Task-Adaptive Multi-Layer Query Fusion Module

$$H_Q = \{Q_0, Q_1, \dots, Q_L\} \# (7)$$

式(7)中, H_Q 表示多层查询状态集合, Q_0 表示初始查询向量, Q_l 表示第 l 层查询向量, L 表示Two-Way Transformer的层数。为了融合这些查询状态, 本文引入一组可学习层级权重参数 $s = \{s_0, s_1, \dots, s_L\}$ 。需要说明的是, 这组权重不是由每个输入样本单独生成的实例级动态权重, 而是在端到端训练过程中学习得到的任务级层级偏好。其归一化形式为:

$$\alpha_l = \frac{\exp(s_l)}{\sum_{k=0}^L \exp(s_k)} \# (8)$$

$$Q_f = \sum_{l=0}^L \alpha_l Q_l \# (9)$$

式(8)和式(9)中, s_l 表示第 l 层查询的可学习权重参数, α_l 表示归一化后的第 l 层查询权重, Q_f 表示融合后的查询向量。随后, 本文使用 Q_f 替代原始末层查询 Q_L , 并将其送入SAM2的最终token-to-image注意力层。设更新后的图像token为 K , 查询位置编码为 P_q , 图像位置编码为 P_i , 则最终查询更新为:

$$Q_{out} = LN(Q_f + \text{Attn}(Q_f + P_q, K + P_i, K)) \# (10)$$

式(10)中, Q_{out} 表示最终输出查询, $LN(\cdot)$ 表示层归一化函数, $\text{Attn}(\cdot)$ 表示注意力函数。该机制既保留了 Q_0 和浅层查询中的定位线索, 也利用了深层查询中经过多轮视觉—语言交互后的语义信息。相比简单拼接, 它不会引入明显维度膨胀; 相比只使用末层查询, 它为最终掩码预测提供了更完整的层级信息。

2.4 文本加权Sobel边界损失

遥感目标边界常呈现狭长、碎片化和低对比度特征。仅依赖区域级损失时, 模型可以获得较好的整体重叠, 但仍可能在道路、建筑边缘和船只轮廓等位置产生模糊或断裂。本文把边界监督与指代表达联系起来, 设计文本加权Sobel边界损失。

$$M_{pred} = \sigma(Z_{fg}) \# (11)$$

$$G(M) = \sqrt{(S_x * M)^2 + (S_y * M)^2 + \varepsilon} \# (12)$$

式(11)和式(12)中, Z_{fg} 表示模型输出 logits 的前景通道, M_{pred} 是连续前景概率图, M 表示待计算边界响应的掩码图, $G(M)$ 表示掩码图的边界响应, S_x 和 S_y 分别表示Sobel水平和垂直卷积核, $*$ 表示卷积运算, ε 取 1×10^{-6} 以避免数值不稳定。对预测掩码和真实掩码分别进行上述计算, 可以得到 G_{pred} 和 G_{gt} 。

为了引入文本条件约束, 模型将文本特征 text_feat 中的句级表示映射为单通道权重, 并扩展到预测掩码尺寸, 得到文本权重图 W_{text} 。该权重图用于同时加权预测掩码和真实掩码的Sobel边界响应, 最终的文本加权边界损失为:

$$L_{bound} = \text{MSE}(G_{pred} \odot W_{text}, G_{gt} \odot W_{text}) \# (13)$$

$$L_{total} = (1 - \lambda)L_{CE} + \lambda L_{Dice} + \mu L_{bound} \# (14)$$

式(13)和式(14)中, L_{bound} 表示文本加权边界损失, $\text{MSE}(\text{mean squared error})$ 表示均方误差函数, $\odot$ 表示逐元素乘法, L_{total} 表示总损失, L_{CE} 表示交叉熵损失, L_{Dice} 表示Dice损失, λ 和 μ 分别表示Dice损失和边界损失的权重。当前代码实现中, λ 取0.1, μ

取0.2。

3 实验及结果分析

3.1 数据集与评价指标

实验使用 RRSIS-D 和 RefSegRS 两个遥感图像指代分割数据集。每个样本由遥感图像、自然语言描述和对应二值掩码组成。RRSIS-D 来源于 RMSIN 工作, 包含 17,402 个图像—文本—掩码三元组 (Liu 等, 2024a); RefSegRS 由 RRSIS 工作提出, 可用于检验自然图像指代分割方法迁移到遥感场景时的表现 (Yuan 等, 2024)。主实验和模块组合消融实验均在 RefSegRS 与 RRSIS-D 两个数据集上进行, 以考察所提模块在不同数据规模、图像尺寸、类别覆盖和目标分布条件下的有效性; 不同边界算子的对比实验则在数据规模和类别数量更大的 RRSIS-D 数据集上开展。

数据集统计: RRSIS-D 包含训练集 12,181、验证集 1,740、测试集 3,481, 输入分辨率为 800×800 , 来源为 (Liu 等, 2024a); RefSegRS 包含训练集 2,172、验证集 431、测试集 1,817, 输入分辨率为 512×512 , 来源为 (Yuan 等, 2024)。

为使实验结果与前文所述科学问题相对应, 本文结合原始遥感图像、自然语言指代表达和真实掩码, 按典型困难特征对 RRSIS-D 与 RefSegRS 测试集样例进行筛选。场景中存在多个外观相近候选区域且需依靠属性、方位或对象关系确定目标的样例归为相似目标干扰; 目标面积较小, 或多个目标在局部区域内分布的样例归为小目标及密集目标; 真实掩码呈细长、弯曲、局部断续或多连通形态, 或目标边界与邻近背景在颜色、纹理上区分度较低的样例归为复杂边界。图 4 和图 5 所示样例至少具有上述一类特征。

评价指标采用平均交并比 (Mean Intersection over Union, mIoU)、总体交并比 (Overall Intersection over Union, oIoU) 以及不同阈值下的 Precision@K (precision at threshold K)。mIoU 先计算每个指代表达样本的预测掩码与真实掩码交并比, 再对样本取平均; oIoU 在整个测试集上累积交集和并集后计算整体交并比; Precision@K 表示 IoU 高于阈值 K 的样本比例, 其中 $K \in \{0.5, 0.6, 0.7, 0.8, 0.9\}$

3.2 实现环境

本实验在 Ubuntu 环境下进行, 硬件平台搭载 28 核 CPU 与 240GB 内存, 并配备两张拥有 32GB 显存的 RTX 5090 GPU。实验采用多 GPU 分布式训练, 每张 GPU 对应一个训练进程, 并结合 DeepSpeed ZeRO Stage 2 与 bf16 混合精度以降低显存占用、提升训练效率。主图像输入尺寸统一为 1024×1024 , 并采用 ImageNet 归一化, 以保留遥感图像中小目标、细长结构和边界细节。沿用基线方法 RS2-SAM2 的训练配置, RRSIS-D 数据集规模较大, 主实验训练 40 轮已达到充分收敛; 而 RefSegRS 数据集规模较小, 为保证模型进行足够次数的参数更新, 其相关实验及消融实验训练 60 轮。所有实验均在训练的最后 10 轮中将学习率衰减为原来的 1/10。对 RS2-SAM2、SDM-SAM2 及各消融设置, 采用相同的数据划分、输入尺寸、批大小和评价脚本。其他已发表方法的结果取自其原论文, 因此仅用于同一公开数据划分下的参考比较, 以保证结果公平可比。

优化器采用 AdamW, 基础学习率为 $5e-5$, 权重衰减为 $1e-2$ 。为适应不同模块的训练需求, 采用分组学习率策略: 预训练骨干及大部分参数使用基础学习率, 视觉—语言融合模块、适配器模块、文本映射层和掩码预测相关模块使用较高学习率 $1e-4$, 以加快任务相关模块的收敛, 同时避免过度扰动预训练表征。

完整模型的损失函数由类别加权交叉熵损失、Dice 损失和文本加权 Sobel 边界损失组成; 在消融实验中, 根据对应设置启用或移除文本加权 Sobel 边界损失, 其余损失项及训练配置保持一致。其中, 类别加权交叉熵用于缓解前景与背景像素比例不均衡, Dice 损失进一步增强区域重叠约束, 文本加权 Sobel 边界损失则强化语言相关目标边界的学习。当前实现中, Dice 损失项权重 $\lambda=0.1$ 沿用基线方法 RS2-SAM2 的设置; 新增的文本加权 Sobel 边界损失项权重 μ 设为 0.2。该设置在保持区域分割损失主导作用的同时, 为目标边界提供辅助约束。Dice 类损失常用于处理强类别不平衡分割问题 (Milletari 等, 2016), 但在训练初期预测交集较小时容易引发梯度震荡 (Yeung 等, 2022), 因此将其权重设为较小的 0.1 作为辅助正则项; 同时, 鉴于图像中目标边界像素极其稀疏, 将边界损失权重设为 0.2 以确保网络对边界细节具备足够的学习灵敏度。

3.3 与现有方法比较

本文在 RefSegRS 和 RRSIS-D 上将 SDM-SAM2 与现有图像指代分割及遥感图像指代分割方法进行比较,包括 RRN(Recurrent Refinement Networks)(Li 等, 2018)、CMSA(Cross-Modal Self-Attention Network)(Ye 等, 2019)、BRINet(Bi-directional Relationship Inferring Network)(Hu 等, 2020)、LSCM(Linguistic Structure Guided Context Modeling)(Hui 等, 2020)、CMPC(Cross-Modal Progressive Comprehension)(Huang 等, 2020)、CMPC+(Cross-Modal Progressive Comprehension+)(Liu 等, 2022)、CARIS(Context-Aware Referring Image Segmentation)(Liu 等, 2023)、RIS-DMMI(Dual Multi-Modal Interaction Network, DMMI)(Hu 等, 2023)、LGCE(language-guided cross-scale enhancement)(Yuan 等, 2024)、CRIS(Wang 等, 2022)、LAVT(Yang 等, 2022)、RMSIN(Liu 等, 2024a)、EVF-SAM(Zhang 等, 2024)、FIANet(Lei 等, 2025)与 RS2-SAM2(Rong 等, 2026)等。其中, RRN、CMSA、BRINet、LSCM、CMPC、CMPC+、CRIS、LAVT、CARIS 和 RIS-DMMI 为自然图像指代分割方法, LGCE、RMSIN、FIANet 和 RS2-SAM2 为遥感图像指代分割方法, EVF-SAM 为基于基础分割模型的文本提示方法。表 1 和表 2 分别展示了在 RefSegRS 和 RRSIS-D 数据集上的对比结果。最优结果以粗体显示, 次优结果以下划线标出, 表中 Val 和 Test 分别表示验证集和测试集。图 4 和图 5 分别给出了 SDM-SAM2 与基线模型 RS2-SAM2 在 RRSIS-D 和 RefSegRS 典型困难样例上的可视化对比, 用于展示相似候选目标干扰、小目标及多目标和复杂边界场景中的分割表现。

在 RefSegRS 数据集上, SDM-SAM2 相比基线模型 RS2-SAM2 在验证集上的 mIoU 和 oIoU 分别提高 3.96 和 3.21 个百分点, Pr@0.5、Pr@0.6、Pr@0.7、Pr@0.8 和 Pr@0.9 分别提高 0.70、0.69、1.62、8.12 和 38.51 个百分点。在测试集上, mIoU 和 oIoU 分别提高 4.15 和 3.29 个百分点, 五个 Precision 指标分别提高 3.64、3.63、6.05、9.08 和 15.85 个百分点, 表明所提方法在不同重叠阈值下均能改善基础模型的分割结果。此外, 在测试集上, SDM-SAM2 的性能也普遍优于现有方法。

在 RRSIS-D 数据集上, SDM-SAM2 相比 RS2-SAM2 在验证集和测试集 mIoU 分别提高 0.43 和

1.77 个百分点, 同时在验证集全部五个精度阈值指标上也取得了稳定的增长: Pr@0.5、Pr@0.6、Pr@0.7、Pr@0.8 和 Pr@0.9 分别提升了 1.15、0.92、0.69、0.92 和 0.35 个百分点。此外, 在测试集上, SDM-SAM2 的性能也普遍优于现有方法。

值得注意的是, RefSegRS 验证集与测试集在 Pr@0.8 和 Pr@0.9 等高阈值指标上存在较明显差异。RefSegRS 验证集包含 431 个样本, 测试集包含 1,817 个样本, 两个子集的样本规模差异较大; 表 1 中多种对比方法也呈现出相似的指标变化, 反映了该数据集不同子集上的共同表现。与此同时, 高阈值指标要求预测掩码与真实掩码具有更高的重叠程度, 对目标边界及局部区域的预测偏差更为敏感, 因此验证集与测试集之间的差异在 Pr@0.8 和 Pr@0.9 上表现得更为明显。尽管如此, SDM-SAM2 在 RefSegRS 测试集上的 Pr@0.5—Pr@0.8、mIoU 和 oIoU 均优于 RS2-SAM2, 体现了所提方法整体分割性能的提升。

在 RRSIS-D 测试集上, SDM-SAM2 相较于 RS2-SAM2 的 mIoU 提高了 1.77 个百分点, Pr@0.5、Pr@0.6、Pr@0.7 和 Pr@0.8 分别提高了 2.39、2.26、2.10 和 0.74 个百分点。两种方法的 oIoU 结果较为接近, 分别为 78.59% 和 78.99%。其中, mIoU 对各样本的 IoU 等权平均, oIoU 则根据整个数据集的像素交并结果计算, 更易受到大面积目标分割结果的影响。因此, mIoU 及多个精度阈值指标的提升表明, SDM-SAM2 的改进主要体现在不同样本上的平均分割质量和目标定位稳定性, 而非依赖少数大面积目标的像素累计优势。值得说明的是, 本文所述的目标定位能力增强主要体现在模型能够依据文本指令更加稳定地发现并定位目标区域, 这一能力由 mIoU 及 Pr@0.5~Pr@0.8 的一致提升得到体现。虽然 SDM-SAM2 在 Pr@0.9 上略低于基线模型, 但两者差异仅为 0.34 个百分点。由于 Pr@0.9 对预测掩码与真实掩码的重叠程度要求极高, 同时受到目标定位和像素级边界拟合质量的共同影响, 因此该指标不能单独作为定位能力的评价依据。结合可视化结果, 所提方法在小目标及复杂边界场景中表现出更完整的目标响应和轮廓预测。

图 4 中, “左下方的船只”“右侧的网球场”和“红色田径场”分别用于展示小目标与方位定位、相似候选目标排除和属性引导的目标判别; 图 5 中, “停车

表 1 在 RefSegRS 数据集上与其他模型的比较

Table 1 Comparison with other models on the RefSegRS dataset

Method	Pr@0.5		Pr@0.6		Pr@0.7		Pr@0.8		Pr@0.9		mIoU		oIoU	
	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test
BRINet	36.86	20.72	35.53	14.26	19.93	9.87	10.66	2.98	2.84	1.14	38.73	31.51	61.59	58.22
LSCM	56.82	31.54	41.24	20.41	21.85	9.51	12.11	5.29	2.51	0.84	40.59	35.54	62.82	61.27
CMPC	46.09	32.36	26.45	14.14	12.76	6.55	7.42	1.76	1.39	0.22	42.08	40.63	63.55	55.39
CMSA	39.24	28.07	38.44	20.25	20.39	12.71	11.79	5.61	1.52	0.83	43.62	41.47	65.84	64.53
RRN	55.43	30.26	42.98	23.01	23.11	14.87	13.72	7.17	2.64	0.98	50.81	41.88	69.24	65.06
EVF-SAM	57.77	35.17	37.59	22.34	16.24	9.36	4.87	2.86	1.86	0.39	46.98	36.64	59.61	55.51
CMPC+	56.84	49.19	37.59	28.31	20.42	15.31	10.67	8.12	2.78	2.55	47.13	43.65	70.62	66.53
CARIS	68.45	45.40	47.10	27.19	25.52	15.08	14.62	8.87	3.71	1.98	54.30	42.66	75.79	69.74
CRIS	53.13	35.77	36.19	24.11	24.36	14.36	11.83	6.38	2.55	1.21	53.74	43.26	72.14	65.87
LAVT	80.97	51.84	58.70	30.27	31.09	17.34	15.55	9.52	4.64	2.09	61.53	47.40	78.50	71.86
RIS-DMMI	86.17	63.89	74.71	44.30	38.05	19.81	18.10	6.49	3.25	1.00	65.72	52.15	74.02	68.58
LGCE	90.72	73.75	86.31	61.14	71.93	39.46	32.95	16.02	10.21	5.45	72.51	59.96	83.56	76.81
RMSIN	93.97	79.20	89.33	65.99	74.25	42.98	29.70	16.51	7.89	3.25	73.84	62.58	82.41	75.72
FIANet	-	84.19	-	77.05	-	61.86	-	33.41	-	7.10	-	68.67	-	78.32
RS2-SAM2	<u>95.36</u>	<u>84.31</u>	<u>94.90</u>	<u>79.42</u>	<u>92.58</u>	<u>70.89</u>	<u>83.76</u>	<u>55.70</u>	<u>36.66</u>	<u>21.19</u>	<u>85.21</u>	<u>73.90</u>	<u>88.03</u>	<u>80.87</u>

附表 1

Method	Pr@0.5		Pr@0.6		Pr@0.7		Pr@0.8		Pr@0.9		mIoU		oIoU	
	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test
SDM-SAM2	96.06	87.95	95.59	83.05	94.20	76.94	91.88	64.78	75.17	37.04	89.17	78.05	91.24	84.16

区域内的重型车辆”“沿道路分布的建筑”和“沿树木分布的自行车道”分别用于展示多个小目标、相似建筑干扰及多区域掩码恢复,以及细长、局部低对比度边界恢复。其中,自行车道目标边界与邻近道路及植被背景在局部颜色和纹理上区分度较低。对比可见,SDM-SAM2在文本指定区域覆盖、非目标区域抑制和轮廓连续性方面更接近真实掩码。

3.4 消融实验

表 3 是 RRSIS-D 上的消融实验设置:Base 表示未加入三个改进模块的基础模型;Sobel 表示文本加权 Sobel 边界损失;QF (task-adaptive multi-layer query fusion) 表示任务自适应多层查询融合;TM (stage-aware text memory) 表示阶段感知文本记忆;

Full 表示同时加入阶段感知文本记忆、任务自适应多层查询融合和文本加权 Sobel 边界损失的完整模型。最优结果以粗体显示,次优结果以下划线标出。

为分析 SDM-SAM2 中不同模块的作用,在 RRSIS-D 测试集上对文本加权 Sobel 边界损失、任务自适应多层查询融合和阶段感知文本记忆进行组合消融。单独引入阶段感知文本记忆模块时,模型取得 67.89% 的 mIoU,较 Base 提高 1.17 个百分点,Pr@0.5 提高 2.30 个百分点,表明在视觉特征更新过程中持续保留目标属性、空间位置 and 对象关系等文本信息,有助于提高文本指定目标的识别与定位能力。单独引入 Sobel 边界损失后,mIoU 由 66.72% 提高至 67.72%,Pr@0.5、Pr@0.6 和 Pr@0.7 均有所提升,说

表 2 在RRSIS-D数据集上与其他模型的比较

Table 2 Comparison with other models on the RRSIS-D dataset

Method	Pr@0.5		Pr@0.6		Pr@0.7		Pr@0.8		Pr@0.9		mIoU		oIoU	
	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test	Val	Test
RRN	51.09	51.07	42.47	42.11	33.04	32.77	20.80	21.57	6.14	6.37	46.06	45.64	66.53	66.43
CMSA	55.68	55.32	48.04	46.45	38.27	37.43	26.55	25.39	9.02	8.15	48.85	48.54	69.68	69.39
LSCM	57.12	56.02	48.04	46.25	37.87	37.70	26.37	25.28	7.93	8.27	50.36	49.92	69.28	69.05
CMPC	57.93	55.83	48.85	47.40	38.50	36.94	25.28	25.45	9.31	9.19	50.41	49.24	70.15	69.22
BRINet	58.79	56.90	49.54	48.77	39.65	39.12	28.21	27.03	9.19	8.73	51.14	49.65	70.73	69.88
CMPC+	59.19	57.65	49.36	47.51	38.67	36.97	25.91	24.33	8.16	7.78	51.41	50.24	70.14	68.64
LGCE	86.10	67.65	60.52	61.53	52.24	51.45	42.24	39.62	23.85	23.33	60.16	59.37	76.68	76.34
RIS-DMMI	70.40	68.74	63.05	60.96	54.14	50.33	41.95	38.38	23.85	21.63	60.72	60.12	77.01	76.20
LAVT	69.54	69.52	63.51	63.63	53.16	53.29	43.97	41.60	24.25	24.94	61.46	61.04	77.59	77.19
EVF-SAM	73.51	72.16	67.87	66.50	58.33	56.59	46.15	43.92	25.92	25.48	64.03	62.75	76.32	76.77
FIANet	-	74.46	-	66.96	-	56.31	-	42.83	-	24.13	-	64.01	-	76.91
RMSIN	74.66	74.26	68.22	67.25	57.41	55.93	45.29	42.55	24.43	24.53	65.10	64.20	78.27	77.79
RS2-SAM2	<u>79.25</u>	<u>77.56</u>	<u>74.08</u>	<u>72.34</u>	<u>63.85</u>	<u>61.76</u>	<u>50.57</u>	<u>47.92</u>	<u>30.40</u>	<u>29.73</u>	<u>68.81</u>	<u>66.72</u>	<u>80.16</u>	<u>78.99</u>
SDM-SAM2	80.40	79.95	75.00	74.60	64.54	63.86	51.49	48.66	30.75	29.39	69.24	68.49	79.47	78.59

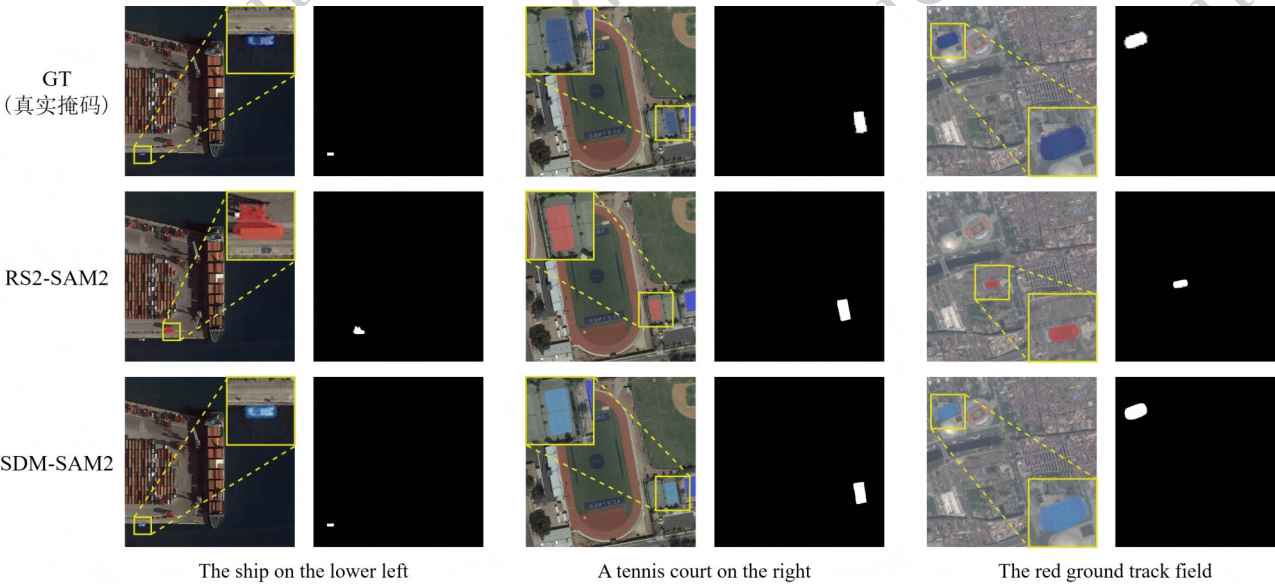


图 4 RRSIS-D数据集上的可视化结果(从左至右:小目标与方位定位、相似目标干扰、属性引导判别)

Fig. 4 Visualization results on the RRSIS-D dataset. (left to right: small-target localization, similar-target interference, and attribute-guided discrimination)

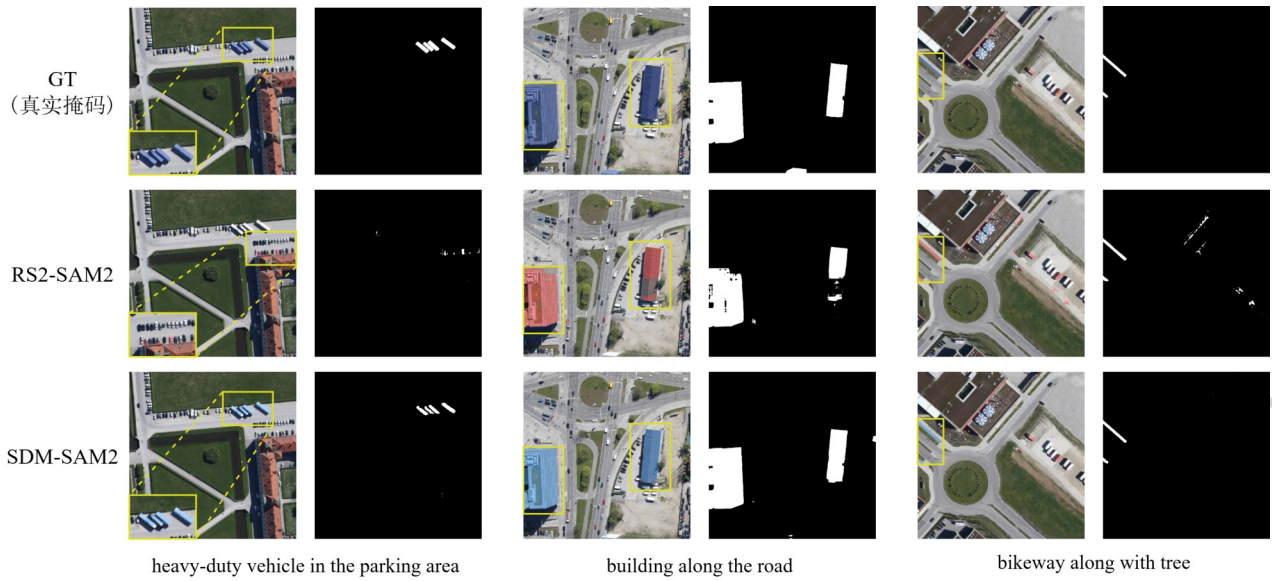


图5 RefSegRS数据集上的可视化结果(从左至右:多个小目标、相似建筑、细长低对比度边界)

Fig. 5 Visualization results on the RefSegRS dataset. (left to right: multiple small targets, similar buildings, and elongated low-contrast boundaries)

明显式边界监督能够改善预测区域与真实掩码的匹配程度。任务自适应多层查询融合同样提升了mIoU,表明不同解码层的查询状态能够为最终掩码预测提供互补信息。

组合实验进一步体现了不同模块之间的功能互补性。Sobel+TM取得67.94%的mIoU和79.10%的oIoU,其中oIoU为所有消融设置中的最高结果,说明文本语义保持与边界约束相结合,有助于改善目标区域的整体重叠质量。QF+TM取得68.40%的mIoU,相较于单独使用TM提高0.51个百分点,并在Pr@0.6、Pr@0.7和Pr@0.8上取得更优结果。这一变化与两个模块的作用位置相对应:阶段感知文本记忆在编码阶段维持指代表达信息,任务自适应多层查询融合在解码阶段整合浅层定位线索和深层语义响应,使文本相关信息能够更充分地参与掩码预测。

完整模型取得68.49%的mIoU,较Base提高1.77个百分点,为所有消融设置中的最高结果;Pr@0.5、Pr@0.6和Pr@0.7分别达到79.95%、74.60%和63.86%,同样取得最优表现。与QF+TM相比,加入文本加权Sobel边界损失后,完整模型的mIoU、oIoU和Pr@0.7分别提高0.09、0.60和0.60个百分点,表明边界监督能够在语义保持和查询融合的基础上进一步优化目标区域预测。mIoU对每个样本

等权计算,主要反映不同尺度目标上的平均分割质量;oIoU由全数据集累计交并比计算,对大面积目标更加敏感,因而不同指标的变化趋势存在一定差异。综合来看,阶段感知文本记忆、任务自适应多层查询融合和文本加权Sobel边界损失分别作用于编码、解码和监督阶段,共同构成从语义保持、信息聚合到边界优化的连续增强链路。

为进一步考察所提出模块在不同数据条件下的适用性,本文在RefSegRS测试集上采用与RRSIS-D相同的模块组合进行了补充消融实验。Base表示未加入三个改进模块的基础模型;Sobel、QF和TM分别表示文本加权Sobel边界损失、任务自适应多层查询融合和阶段感知文本记忆;Full表示同时引入三个模块的完整模型。实验结果如表4所示。

从单模块结果来看,Sobel、QF和TM均能提高Base的mIoU,分别由73.90%提升至77.91%、77.78%和78.47%,说明三个模块分别从边界监督、解码查询聚合和编码阶段语义保持等方面对基础模型形成了有效补充。其中,TM取得78.47%的mIoU和84.34%的oIoU,分别较Base提高4.57和3.47个百分点,为三个单模块中的最优结果;其Pr@0.6、Pr@0.8和Pr@0.9分别提高4.18、8.53和16.34个百分点。这表明,在视觉—语言特征逐层更新过程中保存并检索历史文本状态,有助于维持目标属性、

表3 SDM-SAM2在RRSIS-D测试集上的消融实验

Table 3 Ablation experiments of SDM-SAM2 on the RRSIS-D test set

Method	Sobel	QF	TM	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9	mIoU	oIoU
Base				77.56	72.34	61.76	47.92	29.73	66.72	78.99
Sobel	√			78.91	73.54	62.60	47.97	29.10	67.72	78.47
QF		√		77.85	72.54	61.74	47.46	28.44	66.92	77.40
TM			√	79.86	73.37	62.74	47.74	28.56	67.89	77.54
Sobel+QF	√	√		78.43	73.25	61.76	47.80	28.64	67.11	77.76
Sobel+TM	√		√	79.32	74.26	62.94	48.72	28.90	67.94	79.10
QF+TM		√	√	79.92	74.43	63.26	48.98	29.50	68.40	77.99
Full	√	√	√	79.95	74.60	63.86	48.66	29.39	68.49	78.59

空间位置 and 对象关系等指代信息,进而改善文本指定目标的识别和区域定位。TM在RRSIS-D上同样取得了最高的单模块mIoU,两个数据集呈现出一致趋势,说明阶段感知文本记忆具有较稳定的有效性。

单独引入QF后,模型取得77.78%的mIoU,较Base提高3.88个百分点,Pr@0.5和Pr@0.7分别提高4.41和5.55个百分点,表明不同解码层的查询状态能够为最终掩码预测提供互补信息。单独引入Sobel边界损失后,mIoU和oIoU分别提高4.01和3.43个百分点,同时Pr@0.8和Pr@0.9分别提高8.20和16.45个百分点。由于高阈值Precision要求预测掩码与真实掩码具有更高的区域重叠程度,并且对局部预测偏差更为敏感,该结果支持显式边界监督有助于改善模型的精细分割质量。这与RRSIS-D上Sobel模块提高mIoU及多个Precision指标的结果相互印证。

组合实验呈现出一定的指标相关性。Sobel+QF取得78.17%的mIoU,并在Pr@0.9上达到38.36%,为全部设置中的最高结果,说明查询聚合与边界监督联合使用时,对部分高重叠质量样本具有积极作用。QF+TM同样取得78.17%的mIoU和84.22%的oIoU,较Base分别提高4.27和3.35个百分点,表明编码阶段的语义保持与解码阶段的查询聚合均能够参与改善最终掩码预测。另一方面,Sobel+TM的mIoU为75.98%,虽然仍比Base高2.08个百分点,但低于相应的两个单模块;QF+TM的mIoU也略低于TM单模块。因此,RefSegRS上的组合结果并未呈现模块数量增加、性能随之单调提升的关系,说明不同模块的作用具有评价指标和数据条件相关性。

完整模型取得78.05%的mIoU和84.16%的oIoU,较Base分别提高4.15和3.29个百分点;Pr@0.5、Pr@0.6、Pr@0.7、Pr@0.8和Pr@0.9分别提高3.64、3.63、6.05、9.08和15.85个百分点。其中,完整模型在Pr@0.7和Pr@0.8上分别达到76.94%和64.78%,均为全部消融设置中的最高结果,表明三个模块联合使用时能够提高中高重叠阈值下的分割准确率。与此同时,完整模型的mIoU和oIoU分别比TM单模块低0.42和0.18个百分点,说明其优势主要体现在特定Precision指标,而不是所有评价指标上的一致最优。综合各项结果,完整模型在保持整体区域分割性能的同时,进一步改善了中高重叠阈值下的目标分割质量。

综合两个数据集的消融结果可以发现,三个模块单独加入时均能在RRSIS-D和RefSegRS上提高Base的mIoU,且TM在两个数据集上均取得最高的单模块mIoU,说明三个改进方向具有较好的跨数据集一致性,其中编码阶段维持文本语义约束表现出较稳定的有效性。与此同时,两个数据集上的模块组合呈现出不同的性能变化趋势,这与其主要任务难点不同有关。RRSIS-D包含17,402个图像—文本—掩码三元组,覆盖20个目标类别,图像尺寸为800×800,在数据规模、类别覆盖、目标尺度与方向变化以及场景布局等方面具有更高的多样性,模型需要同时处理语义判别、层级定位和复杂边界等多种问题。因此,阶段感知文本记忆、任务自适应多层查询融合和文本加权Sobel边界损失能够从不同环节形成补充,使完整模型取得最高mIoU。

RefSegRS包含4,420个图像—文本—掩码三元组
© 中国图象图形学报版权所有

组,覆盖 14 个目标类别,图像尺寸为 512×512,其中包含大量小型、分散目标,其主要困难更集中于文本指定目标的识别与定位。在该数据集上, TM 取得最高 mIoU 和 oIoU,说明持续保持文本语义对其主要任务难点具有更直接的改善作用。加入 QF 和 Sobel 后,完整模型在 Pr@0.7 和 Pr@0.8 上取得最优结果,但 mIoU 和 oIoU 分别比 TM 低 0.42 和 0.18 个百分

点,表明不同模块在 RefSegRS 上的增益主要体现在不同评价维度。与此同时,完整模型相较 Base 的 mIoU 和 oIoU 仍分别提高 4.15 和 3.29 个百分点。因此,两个数据集上的结果差异并不表示新增模块失效,而是不同数据分布和主要任务难点使各模块的相对贡献发生了变化。

表 4 SDM-SAM2在 RefSegRS 测试集上的消融实验
Table 4 Ablation experiments of SDM-SAM2 on the RefSegRS test set

Method	Sobel	QF	TM	Pr@0.5	Pr@0.6	Pr@0.7	Pr@0.8	Pr@0.9	mIoU	oIoU
Base				84.31	79.42	70.89	55.70	21.19	73.90	80.87
Sobel	√			87.67	83.10	75.62	63.90	37.64	77.91	84.30
QF		√		88.72	83.27	76.44	63.46	36.49	77.78	83.79
TM			√	88.55	83.60	76.28	64.23	37.53	78.47	84.34
Sobel+QF	√	√		87.89	82.83	76.77	64.45	38.36	78.17	84.23
Sobel+TM	√		√	86.41	81.73	73.64	60.32	33.08	75.98	83.79
QF+TM		√	√	87.73	83.05	76.28	64.28	37.04	78.17	84.22
Full	√	√	√	87.95	83.05	76.94	64.78	37.04	78.05	84.16

在遥感图像指代分割任务中,建筑物、道路、船只等目标通常具有边界细长、形状复杂、尺度变化较大以及背景纹理丰富等特点。对于本文设计的文本加权边界损失,更重要的是边界响应能否准确刻画目标轮廓,为模型提供稳定、连续的边界约束。因此,边界提取算子的选择应充分考虑其与遥感目标边界特性的适配性。综合经典边缘算子的特点,本文最终采用 Sobel 算子作为边界响应提取方式,其原因在于 Sobel 在计算梯度时引入中心加权平滑,对噪声具有较好的抑制能力,同时能够兼顾水平和垂直方向的梯度变化,更适合刻画遥感目标复杂且连续的边界结构。

为进一步验证 Sobel 算子选择的合理性,本文在保持 SDM-SAM2 模型结构、训练策略以及测试流程一致条件下,关闭 TM 和 QF,仅在 Base 上启用边界损失,将文本加权边界损失中的 Sobel 算子分别替换为 Prewitt 和 Laplacian 算子进行对比实验。其中, Prewitt 与 Sobel 均属于一阶梯度算子。Prewitt 在正交方向采用均匀权重,而 Sobel 采用 [1,2] 的中心加权,在保留方向梯度信息的同时具有更强的局部平滑作用。因此,二者的对比用于考察不同平滑权重

对复杂纹理背景下边界监督稳定性的影响。Laplacian 属于二阶微分算子,对局部高频变化更敏感,用于考察二阶边界响应在当前任务中的适用性。通过引入这两类具有代表性的边界响应方式,可以从梯度计算机制和平滑特性两个角度验证不同算子对模型性能的影响。

实验结果如表 5 所示,采用 Sobel 算子时模型取得了最佳整体性能, mIoU 达到 67.72%, 高于采用 Laplacian (66.83%) 和 Prewitt (67.49%) 时的结果;同时,在 Pr@0.6、Pr@0.7、Pr@0.8 和 Pr@0.9 等指标上也整体优于另外两种算子。该结果表明,在遥感图像指代分割任务中,采用具有局部平滑能力的梯度描述方式能够为边界损失提供更加稳定的监督信号,从而有利于复杂目标边界的优化预测。

阶段感知文本记忆模块与包含复杂空间关系或多个相似目标的语义判别需求相对应;任务自适应多层查询融合与小目标和密集目标场景中的定位需求相对应;文本加权 Sobel 边界损失与道路、自行车道、建筑边缘和船只轮廓等细长、低对比度或局部碎片化边界的连续预测需求相对应。

表 5 SDM-SAM2在RRSIS-D测试集上的不同算子的效果对比

Table 5 Comparison of the effects of different operators for SDM-SAM2 on the RRSIS-D test set

Method	Pr@ 0.5	Pr@ 0.6	Pr@ 0.7	Pr@ 0.8	Pr@ 0.9	mIoU	oIoU
Sobel	78.91	73.54	62.60	47.97	29.10	67.72	78.47
Laplacian	78.14	72.91	61.82	47.29	28.90	66.83	76.87
Prewitt	78.91	72.97	62.31	47.66	28.56	67.49	77.43

3.5 复杂度分析

由于本文在基于SAM2构建的面向遥感图像指代分割的SDM-SAM2框架中,引入了阶段感知文本记忆模块和任务自适应多层查询融合模块,并结合了基于Sobel算子的文本加权边界损失进行模型优化。为明确这些新增设计对模型复杂度的具体影响,本节对SDM-SAM2与基线模型RS2-SAM2进行参数复杂度定量对比分析,对比结果如表6所示。其中,EP(Extra Parameters)表示额外参数量,RP(Relative Parameters)表示相对参数比例,ATP(Added Trainable Parameters)表示新增可训练参数量。

表 6 RS2-SAM2与SDM-SAM2的参数复杂度对比
Table 6 Comparison of parameter complexity between RS2-SAM2 and SDM-SAM2

Method	RS2-SAM2	SDM-SAM2
Params(M)	934.81	961.05
EP (M)	-	+26.24
RP	1.00×	1.028×
ATP(M)	-	+26.24

如表6所示,基线模型RS2-SAM2的总参数量为934.81 M。而所提SDM-SAM2的总参数量为961.05 M,仅额外引入了26.24 M的新增参数,仅增加了2.8%的参数开销,单位‘M’表示百万(million)。同时,新增参数主要来自阶段感知文本记忆模块和任务自适应多层查询融合模块,用于增强不同编码阶段的文本语义保持及多层视觉—语言特征交互;文本加权Sobel边界损失不引入额外可训练参数。以上分析表明,SDM-SAM2在仅增加有限参数规模的情况下实现了遥感图像指代分割性能的提升,在模型规模与分割精度之间取得了较好的平衡。

4 结 论

提出语义动态边界增强的遥感图像指代分割框架SDM-SAM2。该框架围绕文本约束由语义理解向目标定位和边界恢复的传递过程,构建由编码阶段语义保持、解码阶段查询聚合和监督阶段边界建模组成的连续增强链路。阶段感知文本记忆模块保存并检索不同编码阶段的文本状态;任务自适应多层查询融合模块利用不同层查询状态的互补信息;文本加权Sobel边界损失对文本指定目标的轮廓施加显式约束。三者分别承担语义保持、信息聚合和边界优化功能,共同服务于文本指定目标的精确分割。

RefSegRS和RRSIS-D数据集上的实验结果验证了SDM-SAM2的有效性。两个数据集上的消融实验表明,三个组成部分均能从编码、解码或监督环节改善基础模型的mIoU。其中,完整模型在RRSIS-D上取得最高mIoU,在RefSegRS上取得最高的Pr@0.7和Pr@0.8,说明不同模块的联合效果会随数据条件和评价指标发生变化,但其相较基础模型的整体改进趋势保持一致。不同边界算子的比较结果显示,Sobel算子在当前实验设置下具有较好的整体表现。可视化结果进一步表明,所提方法能够在相似目标干扰、小目标和复杂边界场景中获得更加准确的目标定位与掩码轮廓。

尽管本文方法在遥感图像指代分割任务中取得了一定改进,仍存在进一步完善空间。一方面,阶段感知文本记忆和多层查询融合带来了更加复杂的跨模态交互过程,后续可探索更加轻量化的跨模态交互方式;另一方面,不同增强模块在联合训练时仍可能存在优化耦合,未来可进一步研究更稳定的多模块协同机制。此外,当前实验仍主要围绕RefSegRS和RRSIS-D两个公开数据集展开,后续可在更多类型的遥感场景、跨区域数据和开放词汇指代表达中进一步验证模型的适用性。

参考文献(References)

Carion N, Massa F, Synnaeve G, et al. 2020. End-to-End Object Detection with Transformers//Proceedings of the European Conference on Computer Vision. Cham: Springer: 213-229 [DOI: 10.1007/978-3-030-58452-8_13]

- Chen K, Liu C, Chen B, et al. 2026. RSRefSeg 2: Decoupling Referring Remote Sensing Image Segmentation with Foundation Models. *IEEE Transactions on Geoscience and Remote Sensing*, 64: 1-20 [DOI: 10.1109/TGRS.2025.3647535]
- Chen K, Zhang J, Liu C, et al. 2025. RSRefSeg: Referring Remote Sensing Image Segmentation with Foundation Models//*Proceedings of IGARSS 2025 - 2025 IEEE International Geoscience and Remote Sensing Symposium*. Brisbane: IEEE: 1070-1074 [DOI: 10.1109/IGARSS55030.2025.11243338]
- Chen L C, Zhu Y, Papandreou G, et al. 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation//*Proceedings of the European Conference on Computer Vision*. Cham: Springer: 801-818 [DOI: 10.1007/978-3-030-01234-2_49]
- Cheng B, Misra I, Schwing A G, et al. 2022. Masked-Attention Mask Transformer for Universal Image Segmentation//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans: IEEE/CVF: 1280-1289 [DOI: 10.1109/CVPR52688.2022.00135]
- Ding H, Liu C, Wang S, et al. 2021. Vision-Language Transformer and Query Generation for Referring Segmentation//*Proceedings of the IEEE/CVF International Conference on Computer Vision*. Montreal: IEEE/CVF: 16301-16310 [DOI: 10.1109/ICCV48922.2021.01601]
- Dong Z, Sun Y Z, Liu T Z, et al. 2025. DiffRIS: Enhancing Referring Remote Sensing Image Segmentation with Pre-trained Text-to-Image Diffusion Models. *Fundamental Research*, in press [DOI: 10.1016/j.fmre.2025.11.017]
- Hu Y, Wang Q, Shao W, et al. 2023. Beyond One-to-One: Rethinking the Referring Image Segmentation//*Proceedings of the IEEE/CVF International Conference on Computer Vision*. Paris: IEEE/CVF: 4067-4077 [DOI: 10.1109/ICCV51070.2023.00376]
- Hu Z, Feng G, Sun J, et al. 2020. Bi-directional Relationship Inferring Network for Referring Image Segmentation//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle: IEEE/CVF: 4424-4433 [DOI: 10.1109/CVPR42600.2020.00448]
- Huang S, Hui T, Liu S, et al. 2020. Referring Image Segmentation via Cross-Modal Progressive Comprehension//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle: IEEE/CVF: 10488-10497 [DOI: 10.1109/CVPR42600.2020.01050]
- Hui T, Liu S, Huang S, et al. 2020. Linguistic Structure Guided Context Modeling for Referring Image Segmentation//*Computer Vision - ECCV 2020: 16th European Conference*. Glasgow: Springer: 59-75 [DOI: 10.1007/978-3-030-58607-2_4]
- Kervadec H, Bouchtiba J, Desrosiers C, et al. 2019. Boundary Loss for Highly Unbalanced Segmentation//*Proceedings of the 2nd Conference on Medical Imaging with Deep Learning*. London: PMLR, 102: 285-296 [DOI: 10.48550/arXiv.1812.07032]
- Kim N, Kim D, Lan C, et al. 2022. ReSTR: Convolution-Free Referring Image Segmentation Using Transformers//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans: IEEE/CVF: 18145-18154 [DOI: 10.1109/CVPR52688.2022.01761]
- Kirillov A, Mintun E, Ravi N, et al. 2023. Segment Anything//*Proceedings of the IEEE/CVF International Conference on Computer Vision*. Paris: IEEE/CVF: 4015-4026 [DOI: 10.1109/ICCV51070.2023.00371]
- Lei S, Xiao X, Zhang T, et al. 2025. Exploring Fine-Grained Image-Text Alignment for Referring Remote Sensing Image Segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 63: 1-11 [DOI: 10.1109/TGRS.2024.3522293]
- Li K, Vosselman G and Yang M Y. 2025. Scale-Wise Bidirectional Alignment Network for Referring Remote Sensing Image Segmentation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 226: 350-363 [DOI: 10.1016/j.isprsjprs.2025.05.014]
- Li L J, He Y, Xie G, et al. 2024. Cross-layer detail perception and group attention-guided semantic segmentation network for remote sensing images. *Journal of Image and Graphics*, 29(05): 1277-1290 (李林娟, 贺赞, 谢刚, 等. 2024. 跨层细节感知和分组注意力引导的遥感图像语义分割. *中国图象图形学报*, 29(05): 1277-1290) [DOI: 10.11834/jig.230653]
- Li R, Li K, Kuo Y C, et al. 2018. Referring Image Segmentation via Recurrent Refinement Networks//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Salt Lake City: IEEE: 5745-5753 [DOI: 10.1109/CVPR.2018.00602]
- Lin T Y, Dollar P, Girshick R, et al. 2017. Feature Pyramid Networks for Object Detection//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Honolulu: IEEE: 2117-2125 [DOI: 10.1109/CVPR.2017.106]
- Liu M, Jiang X and Zhang X. 2025. CADFormer: Fine-Grained Cross-modal Alignment and Decoding Transformer for Referring Remote Sensing Image Segmentation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18: 14557-14569 [DOI: 10.1109/JSTARS.2025.3576595]
- Liu S, Hui T, Huang S, et al. 2022. Cross-Modal Progressive Comprehension for Referring Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9): 4761-4775 [DOI: 10.1109/TPAMI.2021.3079993]
- Liu S, Ma Y, Zhang X, et al. 2024a. Rotated Multi-Scale Interaction Network for Referring Remote Sensing Image Segmentation//*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle: IEEE/CVF: 26658-26668 [DOI: 10.1109/CVPR52733.2024.02517]
- Liu S, Zeng Z, Ren T, et al. 2024b. Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection//*Proceedings of the European Conference on Computer Vision*. Cham: Springer, 102: 285-296 [DOI: 10.48550/arXiv.1812.07032]

- Springer: 38-55 [DOI: 10.1007/978-3-031-72970-6_3]
- Liu S A, Zhang Y, Qiu Z, et al. 2023. CARIS: Context-Aware Referring Image Segmentation//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa: ACM: 779-788 [DOI: 10.1145/3581783.3612117]
- Long J, Shelhamer E and Darrell T. 2015. Fully Convolutional Networks for Semantic Segmentation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE: 3431-3440 [DOI: 10.1109/CVPR.2015.7298965]
- Milletari F, Navab N and Ahmadi S A. 2016. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation//Proceedings of the 4th International Conference on 3D Vision. Stanford: IEEE: 565-571 [DOI: 10.1109/3DV.2016.79]
- Radford A, Kim J W, Hallacy C, et al. 2021. Learning Transferable Visual Models From Natural Language Supervision//Proceedings of the 38th International Conference on Machine Learning. Virtual: PMLR, 139: 8748-8763 [DOI: 10.48550/arXiv.2103.00020]
- Ravi N, Gabeur V, Hu Y T, et al. 2025. SAM 2: Segment Anything in Images and Videos[EB/OL].[2026-05-04].
<https://openreview.net/forum?id=Ha6RTeWMd0> [DOI: 10.48550/arXiv.2408.00714]
- Rong F, Lan M, Zhang Q, et al. 2026. RS2-SAM2: Customized SAM2 for Referring Remote Sensing Image Segmentation. Proceedings of the AAAI Conference on Artificial Intelligence, 40(11): 8751-8759 [DOI: 10.1609/aaai.v40i11.37828]
- Ronneberger O, Fischer P and Brox T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation//Medical Image Computing and Computer-Assisted Intervention. Cham: Springer: 234-241 [DOI: 10.1007/978-3-319-24574-4_28]
- Ryali C, Hu Y T, Bolya D, et al. 2023. Hiera: A Hierarchical Vision Transformer without the Bells-and-Whistles//Proceedings of the 40th International Conference on Machine Learning. Honolulu: PMLR, 202: 29441-29454 [DOI: 10.48550/arXiv.2306.00989]
- Sun Y, Dong Z, Jiang H, et al. 2026. CroBIM-U: Uncertainty-Driven Referring Remote Sensing Image Segmentation. IEEE Transactions on Geoscience and Remote Sensing, 64: 1-1 [DOI: 10.1109/TGRS.2026.3665448]
- Takikawa T, Acuna D, Jampani V, et al. 2019. Gated-SCNN: Gated Shape CNNs for Semantic Segmentation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul: IEEE/CVF: 5229-5238 [DOI: 10.1109/ICCV.2019.00533]
- Vaswani A, Shazeer N, Parmar N, et al. 2017. Attention Is All You Need. Advances in Neural Information Processing Systems, 30: 5998-6008 [DOI: 10.48550/arXiv.1706.03762]
- Wang L A, Miao P H, Su W, et al. 2023. Multimodal referring expression comprehension based on image and text: a review. Journal of Image and Graphics, 28(05): 1308-1325 (王丽安, 缪佩翰, 苏伟, 等. 2023. 图像—文本多模态指代表达理解研究综述. 中国图象图形学报, 28(05): 1308-1325) [DOI: 10.11834/jig.221024]
- Wang W, Bao H, Dong L, et al. 2023. Image as a Foreign Language: BEiT Pretraining for Vision and Vision-Language Tasks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE/CVF: 19175-19186 [DOI: 10.1109/CVPR52729.2023.01838]
- Wang Z, Lu Y, Li Q, et al. 2022. CRIS: CLIP-Driven Referring Image Segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE/CVF: 11686-11695 [DOI: 10.1109/CVPR52688.2022.01139]
- Yang J, Zhang L and Lu H. 2025a. Referring Remote Sensing Image Segmentation with Cross-View Semantics Interaction Network[EB/OL].[2026-05-04].
<https://arxiv.org/abs/2508.01331> [DOI: 10.48550/arXiv.2508.01331]
- Yang Z, Wang J, Tang Y, et al. 2022. LAVT: Language-Aware Vision Transformer for Referring Image Segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE/CVF: 18155-18165 [DOI: 10.1109/CVPR52688.2022.01762]
- Yang Z, Yao H, Tian L, et al. 2025b. A Large-Scale Referring Remote Sensing Image Segmentation Dataset and Benchmark [EB/OL].[2026-05-04].
<https://arxiv.org/abs/2506.03583> [DOI: 10.48550/arXiv.2506.03583]
- Ye L, Roohan M, Liu Z, et al. 2019. Cross-Modal Self-Attention Network for Referring Image Segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE/CVF: 10502-10511 [DOI: 10.1109/CVPR.2019.01075]
- Yeung M, Sala E, Schönlieb C B, et al. 2022. Unified Focal Loss: Generalising Dice and Cross Entropy-Based Losses to Handle Class Imbalanced Medical Image Segmentation. Computerized Medical Imaging and Graphics, 95: 102026 [DOI: 10.1016/j.compmedimag.2021.102026]
- Yuan Z, Mou L, Hua Y, et al. 2024. RRSIS: Referring Remote Sensing Image Segmentation. IEEE Transactions on Geoscience and Remote Sensing, 62: 1-12 [DOI: 10.1109/TGRS.2024.3369720]
- Zhang T, Wen Z, Kong B, et al. 2025. Referring Remote Sensing Image Segmentation via Multi-Scale Spatially-Guided Joint Prediction. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 18: 1-16 [DOI: 10.1109/JSTARS.2025.3638802]
- Zhang Y, Cheng T, Hu R, et al. 2024. EVF-SAM: Early Vision-Language Fusion for Text-Prompted Segment Anything Model[EB/OL].[2026-05-04].
<https://arxiv.org/abs/2406.20076> [DOI: 10.48550/arXiv.2406.20076]
- Zhao S L and Zhang Q. 2023. Progress in multi-modal image semantic

segmentation based on deep learning. Journal of Image and Graphics, 28(11): 3320-3341 (赵什陆, 张强, 2023. 深度学习多模态图像语义分割前沿进展. 中国图象图形学报, 28(11): 3320-3341) [DOI: 10.11834/jig.220451]

Zhu X, Su W, Lu L, et al. 2021. Deformable DETR: Deformable Transformers for End-to-End Object Detection[EB/OL].[2026-05-04]. <https://openreview.net/forum?id=gZ9hCDWe6ke> [DOI: 10.48550/arXiv.2010.04159]

作者简介

王一飞,男,本科生,主要研究方向为计算机视觉。E-mail: 1204910496@qq.com

刘帆,通信作者,女,副教授,硕士生导师,主要研究方向为遥感图像处理和机器学习。E-mail:liufan@tyut.edu.cn

赵联,男,本科生,主要研究方向为计算机视觉。E-mail: 2279028325@qq.com

李雷,男,本科生,主要研究方向为计算机视觉。E-mail: 13038794185@163.com

游书航,男,本科生,主要研究方向为计算机视觉。E-mail: 1544691823@qq.com