软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2024,35(5):2133−2149 [doi: 10.13328/j.cnki.jos.007025]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

面向遥感视觉问答的尺度引导融合推理网络*

赵恩源[1], 宋 宁[1], 聂 婕[1], 王 鑫[2], 郑程予[1], 魏志强[1,3]

[1](中国海洋大学 信息科学与工程学部, 山东 青岛 266100)

[2](清华大学 计算机科学与技术系, 北京 100084)

[3](青岛海洋科技中心, 山东 青岛 266061)

通信作者: 聂婕, E-mail: niejie@ouc.edu.cn; 王鑫, E-mail: xin_wang@Tsinghua.edu.cn

摘 要: 遥感视觉问答 (remote sensing visual question answering, RSVQA) 旨在从遥感图像中抽取科学知识. 近年来, 为了弥合遥感视觉信息与自然语言之间的语义鸿沟, 涌现出许多方法. 但目前方法仅考虑多模态信息的对齐和融合, 既忽略了对遥感图像目标中的多尺度特征及其空间位置信息的深度挖掘, 又缺乏对尺度特征的建模和推理的研究, 导致答案预测不够全面和准确. 针对以上问题, 提出一种多尺度引导的融合推理网络 (multi-scale guided fusion inference network, MGFIN), 旨在增强 RSVQA 系统的视觉空间推理能力. 首先, 设计基于 Swin Transformer 的多尺度视觉表征模块, 对嵌入空间位置信息的多尺度视觉特征进行编码; 其次, 在语言线索的引导下, 使用多尺度关系推理模块以尺度空间为线索学习跨多个尺度的高阶群内对象关系, 并进行空间层次推理; 最后, 设计基于推理的融合模块来弥合多模态语义鸿沟, 在交叉注意力基础上, 通过自监督范式、对比学习方法、图文匹配机制等训练目标来自适应地对齐融合多模态特征, 并辅助预测最终答案. 实验结果表明, 所提模型在两个公共 RSVQA 数据集上具有显著优势.

关键词: 遥感视觉问答; 多模态智能融合; 多模态推理; 多尺度表征

中图法分类号: TP18

中文引用格式: 赵恩源, 宋宁, 聂婕, 王鑫, 郑程予, 魏志强. 面向遥感视觉问答的尺度引导融合推理网络. 软件学报, 2024, 35(5): 2133–2149. http://www.jos.org.cn/1000-9825/7025.htm

英文引用格式: Zhao EY, Song N, Nie J, Wang X, Zheng CY, Wei ZQ. Scale-guided Fusion Inference Network for Remote Sensing Visual Question Answering. Ruan Jian Xue Bao/Journal of Software, 2024, 35(5): 2133–2149 (in Chinese). http://www.jos.org.cn/1000-9825/7025.htm

Scale-guided Fusion Inference Network for Remote Sensing Visual Question Answering

ZHAO En-Yuan[1], SONG Ning[1], NIE Jie[1], WANG Xin[2], ZHENG Cheng-Yu[1], WEI Zhi-Qiang[1,3]

[1](Faculty of Information Science and Engineering, Ocean University of China, Qingdao 266100, China)

[2](Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

[3](Qingdao Marine Science and Technology Center, Qingdao 266061, China)

Abstract: Remote sensing visual question answering (RSVQA) aims to extract scientific knowledge from remote sensing images. In recent years, many methods have emerged to bridge the semantic gap between remote sensing visual information and natural language. However, most of these methods only consider the alignment and fusion of multimodal information, ignoring the deep mining of multi-scale features and their spatial location information in remote sensing image objects and lacking research on modeling and reasoning about scale features,

* 基金项目: 国家重点研发计划 (2021YFF0704000); 国家自然科学基金 (62172376); 国家自然科学基金区域创新发展联合基金 (U22A2068); 中央引导地方科技发展专项资金 (YDZX2022028)
赵恩源和宋宁为共同第一作者.
本文由“多模态协同感知与融合技术”专题特约编辑孙立峰教授、宋新航副研究员、蒋树强教授、王莉莉教授、申恒涛教授推荐.
收稿时间: 2023-04-10; 修改时间: 2023-06-08; 采用时间: 2023-08-23; jos 在线出版时间: 2023-09-11
CNKI 网络首发时间: 2024-01-29

thus resulting in incomplete and inaccurate answer prediction. To address these issues, this study proposes a multi-scale-guided fusion inference network (MGFIN), which aims to enhance the visual spatial reasoning ability of RSVQA systems. First, the study designs a multi-scale visual representation module based on Swin Transformer to encode multi-scale visual features embedded with spatial position information. Second, guided by language clues, the study uses a multi-scale relation reasoning module to learn cross-scale higher-order intra-group object relations with scale space as clues and performs spatial hierarchical inference. Finally, this study designs the inference-based fusion module to bridge the multimodal semantic gap. On the basis of cross-attention, training goals such as self-supervised paradigms, contrastive learning methods, and image-text matching mechanisms are used to adaptively align and fuse multimodal features and assist in predicting the final answer. Experimental results show that the proposed model has significant advantages on two public RSVQA datasets.

Key words: remote sensing visual question answering (RSVQA); multimodal intelligent fusion; multimodal reasoning; multiscale representation

随着深度学习和卫星传感器系统的不断发展, 遥感技术在许多实际应用中扮演着重要角色, 例如灾害监测、农业管理和军事安全[1,2]. 近年来, 遥感图像解读和分析的主流技术包括对象检测[3,4]、场景分类[5−9]、图像匹配[10−12]以及语义分割[13−16]. 以上任务旨在识别图像中的各种目标并提取有用信息. 然而, 以上技术无法捕捉遥感图像中目标之间的视觉关系, 包括空间关系和语义关系, 以上关系依赖于图像中隐含的高层信息. 随着遥感技术的不断发展, 为了更好地理解场景的高级语义信息并学习物体之间的关系, 一些基于语言的视觉理解工作逐渐兴起. 此类任务包括图像描述[17−19]、图像文本检索[20−22]和视觉问答[23−26]. 通过此类任务, 可以深入挖掘遥感图像中隐含的信息, 并实现更精准、更全面的遥感应用. 随着多模态深度学习和自然语言处理技术的发展, 视觉问答 (visual question answering, VQA) 作为一种多模态视觉理解任务, 受到研究者的广泛关注. 在 VQA 任务中, 系统需要基于图像的文本问题推断出答案[27]. VQA 模型包括 3 个基本步骤[28]: 1) 分别为图像和问题构建具有表达能力的表示; 2) 将视觉特征和文本特征进行融合, 生成图像-文本联合表示; 3) 将融合的图像-问题特征输入到多分类器中, 从答案空间中预测最佳匹配答案.

回答有关卫星图像的自然语言问题是智能系统认知能力的体现. 如 Lobry 等人[23]首次将 VQA 系统引入遥感领域. 这项先驱性工作的主要贡献是发布了两个数据集, 其中包括低分辨率和超高分辨率遥感图像及相应的问题答案对以用于各种任务, 例如存在/缺失判断、农村/城市场景分类、目标计数等. 此外, 该工作提出了一种简单的联合嵌入方法, 并探讨遥感数据中 VQA 任务的痛难点. 该方法没有考虑遥感图像的空间信息和图像-问题的交互作用, 仅将图像特征和问题特征简单融合成单个向量, 然后输入到全连接层以预测答案. 而后续工作都致力于图文表征的深度融合, 以挖掘多模态特征的信息价值. 例如, Bazi 等人[25]提出了一种改进方案, 使用注意力机制和双线性技术增强多模态联合表征. 考虑到图像特征和语言信息之间的对齐, 受到人类学习过程的启发, Yuan 等人[24]开发了一种渐进式学习方法, 从易到难地构建训练过程, 以提高模型对知识的认知. Zhang 等人[26]提出一种基于哈希编码的空间多尺度视觉表示模块来处理遥感图像以丰富其空间信息, 并通过空间分层推理模块学习文本引导的内部组视觉和语义关系.

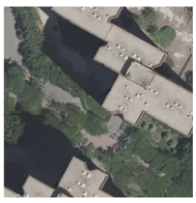
遥感图像与自然图像的视觉特征有很大的差异, 主要体现在地理空间对象的多样性和尺度差异维度. 现以 RSVQA 数据集[23]为例, 如图 1 所示, 其中图 1(a) 中, 文本端蓝色字体所代表的实体, 即图 1(a) 中框选的视觉实体, 在尺度上存在较大差异. 在图 1(b) 中, 许多问题都涉及实体间关系的推理, 如卫星图像中地理空间对象的数量或位置关系 (相邻、右侧、蕴含), 即红色字体标注部分. 尽管既有工作在推动遥感视觉问答研究方面取得了一定的进展, 但现有方法仍然存在以下限制.

Q: 图像中的建筑物是否多于道路?
A: 是

Q: 图中有几辆车停泊在路边?
A: 2

Q: 住宅区旁边是否存在建筑物?
A: 是

(a) 遥感实体尺度差异示例



Q: 路上是否有住宅楼?
A: 是

Q: 道路右侧是否有建筑物?
A: 是

Q: 图中有几栋建筑?
A: 3

(b) 遥感实体位置关系示例

图 1 遥感视觉问答任务示例

(1) 尺度差异导致无效表征. 缺乏对多尺度视觉信息的有效表征, 使得模型在不同尺度下的特征提取和匹配能力不足, 难以准确推断出具有尺度差异的地理空间对象之间的空间关系. 此外, 遥感图像中前景比例远小于自然图像, 因此视觉问答模型还需要解决前景-背景不平衡问题, 并增强对小目标的识别能力.

(2) 视觉空间推理能力不足. 遥感视觉问题的特点之一是卫星图像中包含了大量且分布复杂的地理空间对象, 这就对遥感 VQA 模型提出了更高的视觉空间推理能力的要求. 然而, 现有模型大多采用简单的注意力机制来提取与问题相关的视觉特征, 而忽略了在推理阶段对地理空间对象之间的空间和语义关系进行建模和学习. 这就导致模型难以处理涉及多个尺度、位置、方向、数量等空间信息的问题.

(3) 遥感数据多模态融合效率低下. 融合不充分是视觉问答的领域问题, 即使抽取到有效特征并得到丰富的高阶信息, 但如不能将其映射到共现语义空间内, 则会影响下游的回答质量. 且遥感数据往往会受到各种环境因素如干扰、遮挡、光照变化等的影响, 导致数据质量不稳定和不一致. 为了提高模型的鲁棒性和泛化能力, 需要对不同模态的数据进行有效的融合, 以减轻数据之间的差异和冲突, 并削弱噪声减益. 然而, 现有模型大多采用简单的融合机制来整合与问题相关的视觉特征, 而忽略了对不同模态之间内在联系和相互补充性的挖掘和利用.

针对上述问题, 本文提出了一种多尺度引导的融合推理网络 (multi-scale guided fusion inference network, MGFIN), 以提升 RSVQA 的性能. 通过抽取多尺度特征并对其关系进行统一建模后融入到 RSVQA 框架, 提升模型对遥感场景下多尺度特征及其高阶关系等丰富语义的理解. 具体来说, 在多尺度特征的构建过程中, 本文加入了全局和局部位置编码并通过软硬注意力机制对多尺度特征进行有效抽取, 确保空间信息的精准刻画. 在推理过程中, MGFIN 构建了多尺度特征的空间关系, 有效对多尺度对象间细粒度的空间关系进行丰富表达. 在融合过程中, MGFIN 把推理模块作为桥梁, 通过交叉注意力机制, 又引入对比学习损失, 文本匹配损失, 语言掩码损失等目标函数实现单模态数据流与多模态数据流的对齐和融合. 总之, MGRIN 在问题文本信息的指导下学习遥感多尺度对象的高阶关系, 并收集丰富的关系感知视觉特征, 并进一步学习更强大的图像-问题联合嵌入来预测答案. 本文主要贡献如下.

(1) 提出了一种多尺度引导的融合推理网络, 通过对多尺度关系的推理融合, 获得了更丰富的信息表示.

(2) 建模了多尺度对象间空间关系与文本间的语义关联这两种互补的先验知识, 以文本语义为线索, 实现了模型对多尺度对象间关系的推理.

(3) 构建了融合编码器, 通过在单模态编码器顶层和交叉编码器层加入推理机制, 实现多尺度视觉和文本表示在交叉模态编码器中自底向上地对齐和融合.

(4) 在 RSVQA-LR 数据集和 RSVQA-HR 数据集上进行了充分的对比实验和消融实验. 实验结果表明, MGFIN 与现有最好方法相比具有更出色的表现.

本文第 1 节回顾自然图像和遥感图像视觉问答任务相关工作并提出限制与挑战. 第 2 节刻画 MGRIN 模型细节, 展示子模块, 给出数据流与损失函数. 第 3 节进行实验结果分析. 第 4 节进行总结并讨论未来研究方向.

1 相关工作

1.1 视觉问答

视觉问答 (visual question answering, VQA) 是一类跨学科综合性问题, 涉及计算机视觉 (computer vision, CV) 和自然语言处理 (natural language processing, NLP) 技术, 并在近年来持续发展. 目前, VQA 研究主要集中在多模态联合表示[27,28]和视觉注意机制[25,26]两个方面. 早期工作[23]通常采用简单的逐元素求和/乘积或直接串联来融合多模态特征. 而现有工作则使用更复杂和富有表现力的融合策略, 如多模态紧凑双线性池化 (multimodal compact bilinear pooling, MCB)[29]、多模态低秩双线性注意网络 (multimodal low-rank bilinear attention network, MLB)[30]和多模态分解双线性池化 (multimodal factorized bilinear pooling, MFB)[31]等, 它们利用双线性技术来学习高级别的多模态联合表示. 此外, 许多研究人员也探索了注意机制, 并将其应用于 VQA 模型中, 以增强智能性和可解释性. 例如, Yang 等人[32]提出了多层堆叠注意力网络 (stacked attention networks, SANs), 利用问题中的语义表示作为查询来定位图像中相关的视觉区域; Anderson 等人[33]构建了一种结合了自下而上和自顶向下的注意机制, 以学习

Faster R-CNN[34]检测到的对象级别的图像区域特征; Song 等人[35]提出了立体视觉注意力 (cubic visual attention, CVA), 对问题相关的视觉语义属性进行通道级别的注意选择, 从而进一步丰富图像表示. 然而, 大部分基于注意力的 VQA 方法关注图像中的视觉内容, 而忽略了自然语言问题中隐含的语义信息. 与此不同, 一些协同注意力网络[36]被提出来模拟问题关键词和图像关键对象之间的密集交互.

1.2 遥感域视觉问答

尽管自然图像上的视觉问答 (VQA) 已经取得长足进步, 但在遥感场景上的该任务仍处于起步阶段. Lobry 等人[23]首次将 VQA 系统引入遥感领域, 其主要贡献是发布了两个数据集, 包括低分辨率和高分辨率的遥感图像, 以及相应的问答对, 涵盖了各种任务, 如存在/不存在判断、乡村/城市场景分类、目标计数等. 此外, 该工作采用了一种简单的联合嵌入方法并探讨遥感数据上 VQA 任务的难点. 然而, 该方法没有考虑到图像的空间信息和图像-问题的交互, 仅将图像特征和问题特征简单合并为统一的图文向量, 然后输入至全连接层以预测答案. 随后, 受人类学习过程的启发, Yuan 等人[24]开发了一种渐进式 VQA 学习方法, 按照由易到难的回答逻辑, 调节问题-答案对的难度以训练模型. 与依赖于视觉和文本信息的联合表示不同, Chappuis 等人[37]提出了一种名为 Prompt-RSVQA 的方法, 将视觉信息翻译成单词, 然后注入到仅包含语言的模型中. 最近, Bazi 等人[25]提出了一种针对遥感图像的 VQA 方法, 利用视觉-语言 Transformer 作为图像和问题的编码器, 并通过协同注意力机制建模跨模态依赖性. 值得注意的是, 提取地表覆盖变化信息一直是遥感图像理解的焦点. 最近, 引入了一项新颖而有意义的任务: 基于多时相航拍图像的变化检测视觉问答 (change detection visual question answering, CDVQA)[38], 为遥感场景中的 VQA 任务提供了一条有价值的新研究方向. Zhang 等人[26]提出了基于哈希的空间多尺度视觉表征模块来弥合遥感图像的尺度差异和空间位置敏感性造成的语义鸿沟, 并通过空间分层推理模块学习文本引导的内部组视觉和语义关系.

1.3 关系推理网络

基于实体及其属性之间关系的推理能力是智能系统的关键能力. 近年来, 关系网络作为一种通用的解决方案被广泛应用于各种依赖于关系推理的任务[39−43], 特别是视觉问答任务. 具体而言, 文献 [39] 提出了一种简单而有效的神经网络模块——关系网络, 它可以隐式地推理实体及其关系, 并在视觉问答任务上成功地超越了人类水平. 受到关系推理的启发, Zhou 等人[40]设计了时序关系网络, 它可以在多个时间尺度上进行时序关系推理, 用于活动识别任务. 文献 [41] 提出了一种分层条件架构, 用于视频问答任务, 它允许以阶段性的方式进行高阶关系推理. 随后, Hu 等人[42]提出了一种目标检测网络, 它通过物体外观和几何特征间的特征交互来建模物体之间的关系. 文献 [43] 对航空影像上的语义分割性能进行了改进, 建模并强化了上下文空间关系和通道关系. 以上工作都从一种统一的视角审视, 尽皆采用了关系推理来增强特征表示. 对于遥感视觉问答任务而言, 地理空间对象之间的关系是主要问题. 因此, 在遥感视觉问答模型中获得一种增强了关系信息的视觉表示是必要的. 受此启发, 在本文中, MGFIN 专注于在文本信息的指导下学习推理多尺度对象之间的高阶关系, 以获得丰富的关系感知视觉特征.

2 多尺度引导的融合推理网络

针对遥感视觉问答任务, 本文提出了多尺度引导的融合推理网络 (MGFIN), 通过对多尺度空间关系的建模、推理与融合, 构建知识表征增强的视觉问答模型.

2.1 MGFIN 模型概述

给定图像表征 $\mathcal{V}$ 和相应的问题 $\mathcal{Q}$, VQA 任务的目标是从答案空间 $\mathcal{A}$ 中推断出正确的答案 $\tilde{a}$, 该空间是针对开放式问题的一组预定义候选答案. 通常, 它可以表述如下:

$$\tilde{a}=\underset{a\in\mathcal{A}}{\operatorname{argmax}} f_\theta(a|\mathcal{Q},\mathcal{V}) \tag{1}$$

其中, f 是具有可训练参数 θ 的模型的评分函数, 并通过交叉熵得到损失函数记作 $\mathcal{L}_{\text{vqa}}$.

MGFIN 模型同样以问题和图像两种模态的数据作为输入, 在进行多尺度推理得到多模态数据统一表征后, 输入到多层编码器网络对多模态信息深度融合, 最后输入答案空间进行预测. 其整体架构如图 2 所示, MGFIN 分别

对文本和遥感图像进行向量表征, 经多尺度空间筛选模块滤波清洗后, 再进行多尺度关系推理, 随后经过跨模态桥接融合编码器, 在多种损失函数监督下进行多模态信息的对齐融合.

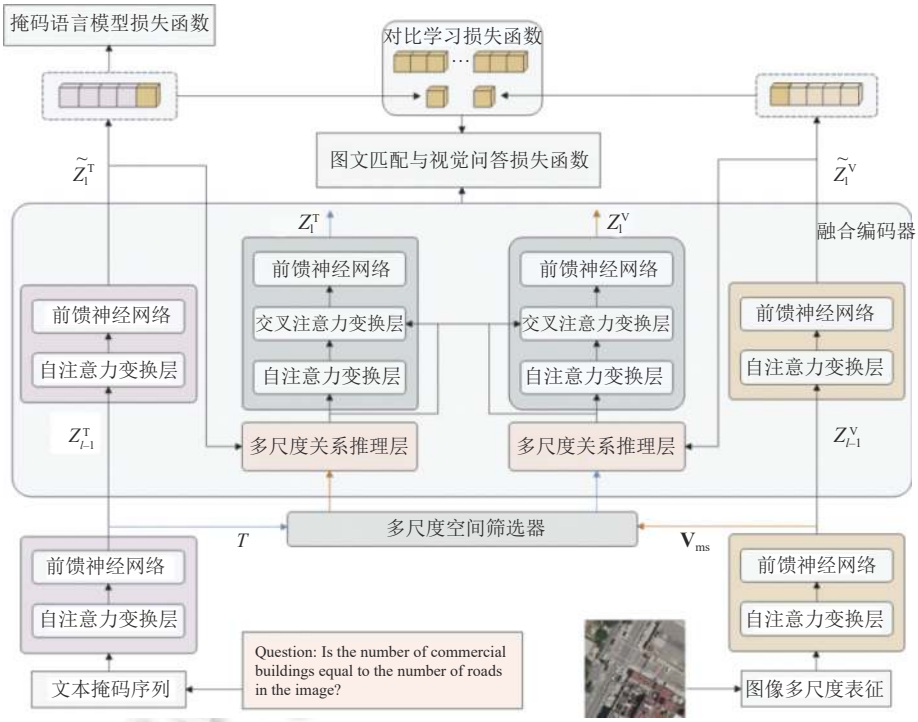


图 2 MGFIN 整体框架图

2.2 多尺度语义表征

2.2.1 多尺度绝对空间特征提取

如图 3 所示, MGFIN 为了进行多尺度特征提取, 采用了 Swin Transformer[44]作为视觉编码器. Swin Transformer 是一种基于滑动窗口自注意力机制的 vision Transformer (ViT), 能够有效地捕捉图像中不同尺度的特征. 它的核心思想是将输入图像划分为若干像素块, 并在每个滑动窗口内进行局部自注意力运算, 从而建模局部依赖关系. 同时, 它还利用图像块合并操作, 将相邻的小块合并成更大的小块, 实现跨尺度特征的提取和融合. 不同尺度的图像块在特征提取上具有互补性; 较大的图像块可以更好地表达粗粒度特征, 较小的区块可以更好地细化细粒度特征. 具体来说, MGFIN 通过图像块分割将图像块大小初始化为 4, 并在多个尺度层次上进行特征提取. 通过图像块合并操作, 分别得到了 $H/8\times H/8$、$H/16\times H/16$ 和 $H/32\times H/32$ 分辨率的多尺度特征. 对于每个输入 2D 图像 $\mathbf{I}\in\mathbb{R}^{H\times W\times C}$, 其中 (H, W) 表示输入图像的分辨率, C 表示通道数, ViT 将其重塑为扁平化的 2D patch 序列 $\mathbf{P}\in\mathbb{R}^{N\times(P^2C)}$, 其中 (P,P) 表示图像 patch 的分辨率, $N=\dfrac{HW}{P^2}$ 表示 patch 的数量. 与 BERT 类似, 本文通过在 Swin Transformer 图像块序列前加入了 [class] 标记, 并使用可学习的 1D 位置嵌入 $\mathbf{V}^{\text{pos}}\in\mathbb{R}^{(N+1)\times D_v}$ 来增强位置信息, 其中 D_v 表示视觉编码器的维度. 输入图像经过视觉编码器后得到如下表示:

$$\mathbf{V}_{\text{ms}}=\left[\mathbf{E}_{[\text{class}]};\mathbf{p}_1\mathbf{W}_{\mathbf{p}}^1;\ldots;\mathbf{p}_N\mathbf{W}_{\mathbf{p}}^i\right]+\mathbf{V}^{\text{pos}} \tag{2}$$

其中, $\mathbf{W}_{\mathbf{p}}^i\in\mathbb{R}^{(P^2C)\times D_v}$ 是可训练的线性投影层, $\mathbf{W}_{\mathbf{p}}^i\in\mathbb{R}^{(P^2C)\times D_v}$. 每一层的 Swin Transformer 块由基于移动窗口的多头自注意力 (SW-MSA) 块和前馈网络 (FFN) 模块构成. 为方便起见, 本文把抽取的多尺度特征 ($H/8\times H/8$、$H/16\times H/16$ 和 $H/32\times H/32$) 统一表示为 $\mathbf{V}_{\text{ms}}$, 并将其简化为 Encoder^V. 第 ℓ 层表示为 $\mathbf{V}_{\text{ms}\ell}=\text{Encoder}_\ell^V(\mathbf{V}_{\ell-1}),\ell=1,\ldots,L_{\text{V}}$, 其中 L_{V} 是视觉编码器的层数.

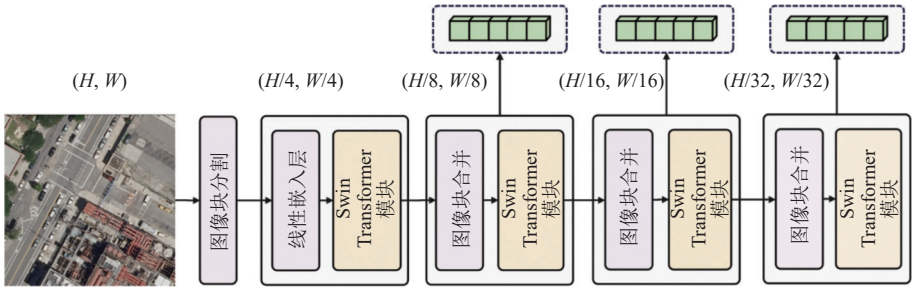


图 3 多尺度表征流程示意图

2.2.2 多尺度相对空间位置嵌入

在 Swin Transformer 中, 滑动窗口机制可以解释为一种局部到全局 (local-to-global) 的策略, 它通过逐步扩大感受野来捕捉图像中不同尺度和位置的信息, 有了这种移位的窗口划分机制, SW-MSA 和 MLP 模块的输出可以写成:

$$\hat{\mathbf{z}}^{l+1} = \text{SW-MSA}\left(\text{LN}\left(\mathbf{z}^{l}\right)\right) + \mathbf{z}^{l} \tag{3}$$

$$\mathbf{z}^{l+1} = \text{MLP}\left(\text{LN}\left(\hat{\mathbf{z}}^{l+1}\right)\right) + \hat{\mathbf{z}}^{l+1} \tag{4}$$

本文使用 $\hat{\mathbf{z}}^{l+1}$ 表示第 $l+1$ 层 Swin Transformer 块中窗口自注意力层输出结果; $\mathbf{z}^{l}$ 表示第 l 层 Swin Transformer 块输出结果. 为了对多尺度相对空间位置进行建模, 本文引入了相对位置偏置这一概念. 该概念指一种能够提升滑动窗口自注意力机制中位置感知能力的技术, 在此基础上可捕获窗口内部及窗口之间存在的相对位置关系. 在计算窗口注意力时, 在参考文献 [1,32,33] 所述方法基础上, 在计算 Q、K 之间相似度时加入了可训练参数 $B \in \mathbb{R}^{M^2 \times M^2}$, 其中 B 代表相对位置的索引:

$$Attention(Q,K,V) = Softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d}} + B\right)V \tag{5}$$

2.2.3 文本表征

本文选用 $\text{BERT}_{\text{BASE}}$ 作为文本编码器, 其在广泛 NLP 任务上具有较高性能. 首先将输入序列 w 通过字节级 BytePair 编码进行分词处理, 并在序列首尾分别加入 [<s>] token 与 [</s>] token 作为起始与结束标记; 然后将输入序列转换成如下形式:

$$T = [\mathbf{E}_{[<s>]}; \mathbf{E}_{w_1}; \ldots; \mathbf{E}_{w_M}; \mathbf{E}_{[</s>]}] + T^{\text{pos}} \tag{6}$$

其中, $T \in \mathbb{R}^{(M+2)\times D_t}$ 为词嵌入矩阵, M 为是分词后序列长度, D_t 为文本编码器的维数, T^{pos} 为位置嵌入矩阵. 接着将该输入送入 $\text{BERT}_{\text{BASE}}$ 模型中进行处理, 文本编码器的第 ℓ 层表示为 $\text{Encoder}_{\ell}^{\mathrm{T}}$, ℓ 层表示为 $T_{\ell} = \text{Encoder}_{\ell}^{\mathrm{T}}(T_{\ell-1}), \ell = 1,\ldots,L_{\mathrm{T}}$, 其中 L_{T} 是文本编码器的层数.

2.3 多尺度空间层次推理融合模块

2.3.1 多尺度空间筛选模块

MGFIN 旨在为多尺度对象关系建模并进行推理, 以挖掘多尺度信息, 获得鲁棒且丰富的表征. 在进行推理前, 为降低计算量并充分提取多尺度特征, 本文首先建立软硬注意力结合的筛选机制, 旨在清除冗余特征. 筛选模块由软注意力部分和硬注意力部分组成. 软注意力部分负责评估不同区域特征的信息权重, 并筛选出重要区域. 硬注意力部分则根据软注意力部分得到的权重, 在重要区域中进一步选择信息. 这样做既可以减少冗余区域带来的计算量, 又可以根据输入问题对不同区域特征赋予不同权重. 本文将软注意力和硬注意力机制结合起来进行多尺度关系推理, 以实现更高效和准确地回答视觉问题.

为了实现上述目标, MGFIN 首先从输入图像经过视觉编码器得到的视觉向量 $\mathbf{V}_{\text{ms}}$ 中提取 Q 个区域特征 $\mathbf{V}_{\text{ms}} = \{\mathbf{V}_{\text{ms}i} | i=1,\ldots,Q\}$. 然后 MGFIN 利用问题特征 T 和区域特征 $\mathbf{V}_{\text{ms}}$ 作为输入, 通过软注意力部分计算每个区域特征的信息权重向量 α:

$$\alpha = Softmax(W_2 \times (ReLU(W_1 \times ([\mathbf{V}_{ms}, T]) + b))) \tag{7}$$

其中, $[\cdot]$ 为串联运算, 矩阵 W_1 和 W_2 分别代表线性层和非线性层中的权重参数, b 为偏差向量, $ReLU(\cdot)$ 和 $Softmax(\cdot)$ 分别代表 *ReLU* 和 *Softmax* 函数. 在确定权重向量 α 后, 计算加权区域特征 $\hat{v} = \alpha \cdot \mathbf{V}_{ms}$. 其中, $\cdot$ 为元素乘法. 值得注意的是, $\hat{v}$ 将用于进一步的全局关系推理. 同时, MGFIN 使用提出的硬注意力机制在 $\mathbf{V}_{ms}$ 中选择权重值最高的 K 个元素, 形成与问题相关的区域特征集合 $\tilde{v}$:

$$\tilde{v} = f_s(\mathbf{V}_{ms}, K) \tag{8}$$

其中, K 是实验中的超参数, 函数 $f_s(\mathbf{V}_{ms}, K)$ 指的是对 $\mathbf{V}_{ms}$ 进行降序排序, 选择值最大的 K 个特征向量构建输出特征集的操作. 值得注意的是, v 将用于不同尺度的局部关系推理, 而其他特征向量在后期处理中被淘汰. 如图 4 所示.

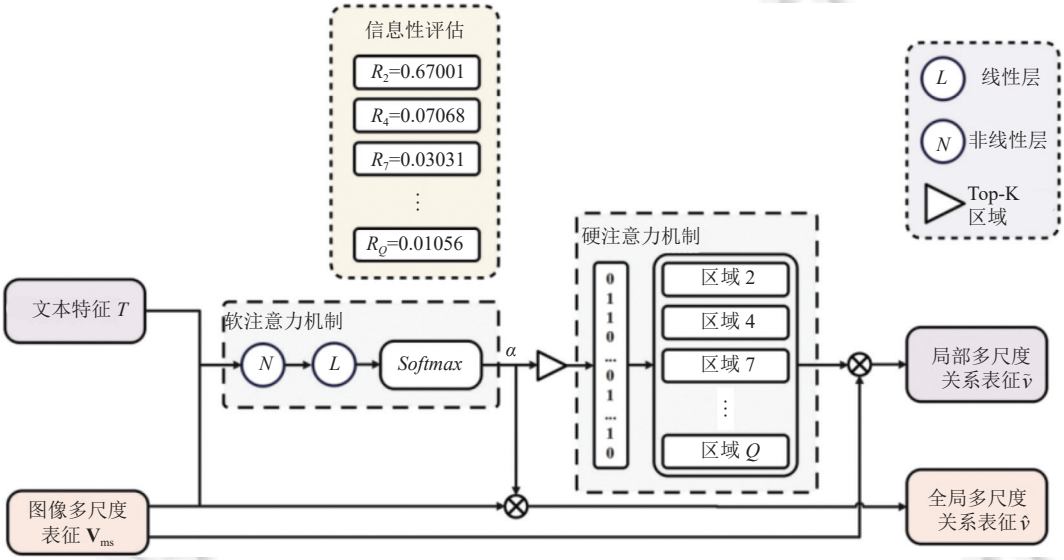


图 4 多尺度空间筛选模块示意图

2.3.2 多尺度关系推理

全局视角利用整幅图像的信息隐式地回答问题; 局部视角模拟多个目标间的关系显式地获得答案. 本文认为这两种模式从不同维度分析视觉信息, 构成了本文提出的关系推理体系结构的基础. 为了适应不同问题带来的挑战, 视觉问答需要在多尺度上进行关系推理, 以实现全面准确地回答问题. 也就是说, 问题的动态变化要求局部关系推理方案预先生成多个目标间的关系描述符, 以降低运行时间和提高反馈速度. 如何进行目标间的关系推理是视觉问答领域广泛讨论的问题, 其常见思路是通过神经网络构造函数描述关系. MGFIN 提出在多尺度上生成尽可能多的关系描述符, 而不是动态地建模关系, 从而解决了目标不稳定带来的复杂性, 并隐式地增强了对不同目标间关系的推理能力. 综上所述, 本文提出的关系推理体系结构基于全局和局部关系推理方案, 并将其输出特征定义为:

$$\mathbf{V}_k = f_g(\hat{v}) + f_l(\tilde{v}, T) \tag{9}$$

其中, T 为输入问题构造的特征, 如公式 (6) 所示. $\tilde{v}$ 和 $\hat{v}$ 分别为与问题相关和加权后的区域特征集合, 函数 $f_g(\hat{v})$ 和 $f_l(\tilde{v}, T)$ 分别代表全局关系推理和局部关系推理方案. 具体来说, 全局关系推理方案先对所有加权区域特征求和, 再通过非线性层计算特征表示. 该非线性层可表示为:

$$f_g(\hat{v}) = ReLU\left(W_g \times \left(\sum_{i=0}^{K} \hat{v}_i\right) + b_g\right) \tag{10}$$

为了更好地解释局部关系推理方案的工作原理, 本文在图 5 中给出了第 3 个尺度上的关系推理示例. 在这个示例中, 模型首先通过区域注意模块提取问题相关区域, 然后将尺度指标定义为一类组合中包含问题相关区域数量. 实验发现, 在 3 个区域和问题特征构成一种组合时效果最佳. 此外, 本文只计算了 M 个区域组合 (其中 M 是一

种超参数), 而不是所有可能的组合, 从而大大节省了计算成本. 将尺度数量 S 定义为用户需要确定的超参数, 所提出的总尺度为 S 的局部关系推理方案可表示为:

$$f_l(\tilde{v},T_s)=R_1(\tilde{v},T_1)+R_2(\tilde{v},T_2)+\ldots+R_s(\tilde{v},T_s) \tag{11}$$

其中, 函数 $R_s(\tilde{v},T_s)$ 表示第 s 个尺度上的关系推理. 本质上, 第 s 个尺度上的关系推理可以形成 $N=C_K^s$ 个组合, 其中 C 表示组合函数. 为了构建高阶关系组合, MGFIN 采用随机选择方法将问题相关区域分成 M 个组合 (其中 M 小于 N). 因此, 第 s 个尺度上关系推理的最终输出可以表示为:

$$R_s(\tilde{v},T_s)=r(c_1,T_1)+r(c_2,T_2)+\ldots+r(c_M,T_s) \tag{12}$$

其中, c_m 表示第 s 个尺度上第 i 个可能的组合, 每个关系项用函数 $r(c_m,T_s)$ 表示, 它捕捉了 m 个有序区域之间的关系, 并且由一层非线性层和一层线性层构成, 用于提取区域之间的关系:

$$r(c_m,T_s)=W_{c,2}\times(ReLU(W_{c,1}\times([c_m,T_s])+b_{c,1}))+b_{c,2} \tag{13}$$

其中, $W_{c,2},W_{c,1}$ 为参数矩阵, $b_{c,1},b_{c,2}$ 为偏置向量.

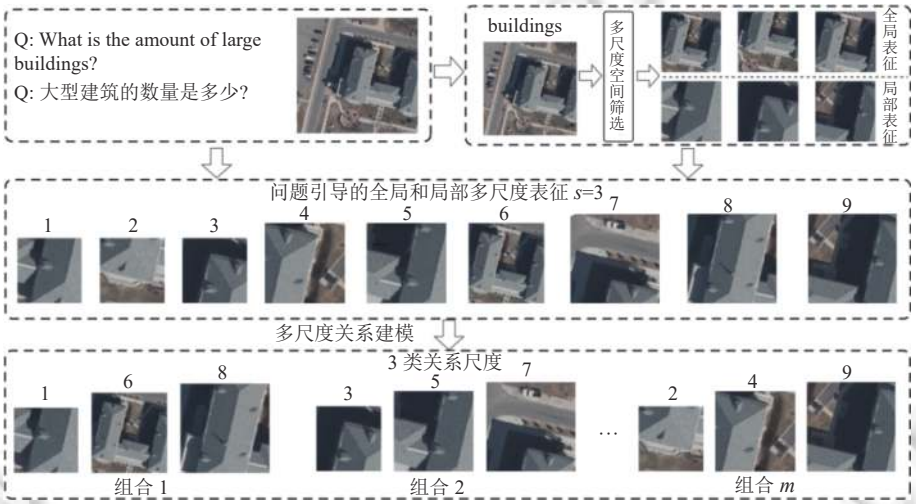


图 5 多尺度关系推理流程示例图

2.3.3 跨模态桥接融合编码器

针对遥感视觉问答任务, 还需要一种良好的融合机制以挖掘丰富的尺度关系表征. 本文受 BridgeTower[45]模型启发, 将多尺度关系推理模块作为模态间交互的桥梁 (桥接层). 具体来说, 桥接层在单模态编码器顶层和跨模态编码器各层之间建立连接. 这样可以在跨模态编码器与单模态编码器的有效自下而上的跨模态对齐和融合形式, 本文将跨模态编码器的第 ℓ 层定义为 Encoder_ℓ^Z, 它由视觉部分和文本部分组成. 每个部分包含多头自注意力块、多头交叉注意力块和一层前馈神经网络. 为了简洁起见, 本文将每层之间的交互定义为:

$$\tilde{\mathbf{Z}}_\ell^{\mathrm{V}}=\mathbf{Z}_{\ell-1}^{\mathrm{V}} \tag{14}$$

$$\tilde{\mathbf{Z}}_\ell^{\mathrm{T}}=\mathbf{Z}_{\ell-1}^{\mathrm{T}} \tag{15}$$

$$\mathbf{Z}_\ell^{\mathrm{V}},\mathbf{Z}_\ell^{\mathrm{T}}=\text{Encoder}_\ell^Z\left(\tilde{\mathbf{Z}}_\ell^{\mathrm{V}},\tilde{\mathbf{Z}}_\ell^{\mathrm{T}}\right),\ell=1,\ldots,L_Z \tag{16}$$

其中, $\mathbf{Z}_\ell^{\{\mathrm{V,T}\}}$ 是第 ℓ 层单模态编码器中视觉或文本部分的输出表示, $\tilde{\mathbf{Z}}_\ell^{\{\mathrm{V,T}\}}$ 是每个部分的输入, L_Z 是跨模态编码器的层数. 当前的视觉语言模型, 直接将前一层的输出表示作为 Encoder_ℓ^Z 的输入 (公式 (14) 和公式 (15)). $\mathbf{Z}_0^{\mathrm{V}}$, $\mathbf{Z}_0^{\mathrm{T}}$ 是单模态编码器得到的最后一层表示进行初始化: $\mathbf{Z}_0^{\mathrm{V}}=\mathbf{V}_{L_{\mathrm{V}}}\mathbf{W}_{\mathrm{V}}+\mathbf{V}^{\text{type}}$, $\mathbf{Z}_0^{\mathrm{T}}=\mathbf{T}_{L_{\mathrm{T}}}\mathbf{W}_{\mathrm{T}}+\mathbf{T}^{\text{type}}$, 其中 $\mathbf{W}_{\mathrm{V}}\in\mathbb{R}^{D_V\times D_Z}$ 和 $\mathbf{W}_{\mathrm{T}}\in\mathbb{R}^{D_T\times D_Z}$ 是线性投影, $\mathbf{V}^{\text{type}}$ 和 $\mathbf{T}^{\text{type}}$ 是模态类型嵌入. 在本文中, $L_{\mathrm{V}}=L_{\mathrm{T}}=12$, $L_Z=6$. 本文提出使用多个桥接层来连接单模态编码器顶层与跨模态编码器每一层:

$$\tilde{\mathbf{Z}}_{\ell}^{\mathrm{V}}=\mathrm{BridgeLayer}_{\ell}^{\mathrm{V}}\left(\mathbf{Z}_{\ell-1}^{\mathrm{V}},\mathbf{T}_{k}\mathbf{W}_{\mathrm{T}}+\mathbf{T}^{\mathrm{type}}\right) \tag{17}$$

$$\tilde{\mathbf{Z}}_{\ell}^{\mathrm{T}}=\mathrm{BridgeLayer}_{\ell}^{\mathrm{T}}\left(\mathbf{Z}_{\ell-1}^{\mathrm{T}},\mathbf{V}_{k}\mathbf{W}_{\mathrm{V}}+\mathbf{V}^{\mathrm{type}}\right) \tag{18}$$

在深度学习中, 层归一化 (LayerNorm) 是一种常见的归一化方法, 可用于调节神经网络层内部的输出和梯度. 它可以使得每个神经元的输出都具有相同的统计特性, 从而使得神经网络的训练更加稳定. 受其启发, 本文利用提出的桥接层将单模态编码器顶层表征与跨模态编码器的每一层相连接, 从而将不同语义层次的单模态表示融入到跨模态交互中. 而第 2.3.2 节提出的多尺度推理桥桥接层本质上就是一种多尺度特征与文本特征间关系的一种跨模态表示, 以各种非线性层的排布, 来对尺度关系进行建模. 而各种非线性层的隐式建模实际上也符合跨模态交互融合极致交互的思想. 因此 MGFIN 采用多尺度关系推理桥作为跨模态编码器与单模态编码器的桥梁, 其简单的形式定义如下:

$$\mathrm{BridgeLayer}_{\mathrm{1st}}(\mathbf{V}_{\mathrm{ms}},T)=f_{g}(\hat{v})+f_{l}(\tilde{v},T) \tag{19}$$

$$\mathrm{BridgeLayer}_{\mathrm{oth}}(\mathbf{V}_{\mathrm{ms}},T)=\mathrm{LayerNorm}(\mathbf{V}_{\mathrm{ms}}+T) \tag{20}$$

其中, $\mathbf{V}_{\mathrm{ms}}$ 代表多尺度视觉表征, T 代表文本表征. 公式 (19) 表示关系推理层的第 1 层, 具体算子含义参考公式 (9), 公式 (20) 代表其他层的融合过程, 两者为递进关系. 第 1 层本质上就是一种多尺度特征与文本特征间关系的一种跨模态表示, 属于单模态数据流的融合. 而其他层则是单模态数据流与多模态数据流的一种交互融合. 简言之, 公式 (19) 在第 1 层单模态交互得到多模态表征, 参与公式 (20) 后续的多模态融合推理.

2.4 损失函数

本文提出了一种多尺度推理融合视觉问答模型, 该模型能够有效利用多尺度推理来提高问题和答案之间的一致性和可信度. 然而, 这种机制也会增加模型结构和参数的复杂度, 降低模型训练和测试时的效率和稳定性, 并可能引入一些错误或不相关的知识或假设. 为解决以上问题, 受 ALBEF[46]预训练过程启发, MGFIN 采用了 4 个损失函数约束来保证多模态信息的语义空间一致性, 并增强图像和文本之间的语义相似度的衡量能力. 除了传统视觉问答损失函数, 本文还使用了图像文本对比学、掩码语言建模和图像文本匹配损失作用于多模态信息的表征与融合.

2.4.1 掩码语言模型损失

掩码语言模型 (masked language modeling, MLM) 利用图像和上下文文本共同预测被掩码的词语. MGFIN 以 15% 的概率随机地对输入词语进行掩码, 并用特殊符号<MASK> 替换被掩码的词语. 预测过程既依赖于周围的文本信息, 也依赖于被掩码的图像特征. MLM 最小化交叉熵损失函数可定义为:

$$\mathcal{L}_{\mathrm{mlm}}=\mathbb{E}_{(v,\widehat{T})}\mathrm{H}\left(y^{\mathrm{msk}},p^{\mathrm{msk}}\left(\tilde{\mathbf{Z}}_{\ell}^{\mathrm{V}},\widehat{\tilde{\mathbf{Z}}_{\ell}^{\mathrm{T}}}\right)\right) \tag{21}$$

其中, $\widehat{\tilde{\mathbf{Z}}_{\ell}^{\mathrm{T}}}$ 是被掩码的文本标记, $\tilde{\mathbf{Z}}_{\ell}^{\mathrm{V}}$ 是多模态编码器输出的视觉表征, $p^{\mathrm{msk}}\left(\tilde{\mathbf{Z}}_{\ell}^{\mathrm{V}},\widehat{\tilde{\mathbf{Z}}_{\ell}^{\mathrm{T}}}\right)$ 是模型预测结果, y^{msk} 是被遮盖的文本标记的真实值.

2.4.2 图像文本匹配损失

图文匹配 (image-text matching, ITM) 预测一对图像和文本是否为正例 (匹配) 或负例 (不匹配). MGFIN 使用多模态编码器的 [CLS] 词语的输出嵌入作为图文对的联合表示, 并在其后添加一层全连接层和 *Softmax* 层, 来预测两类概率 p^{itm} . ITM 损失函数为:

$$\mathcal{L}_{\mathrm{itm}}=\mathbb{E}_{(v,T)}\mathrm{H}\left(y^{\mathrm{itm}},p^{\mathrm{itm}}\left(\tilde{\mathbf{Z}}_{\ell}^{\mathrm{V}},\tilde{\mathbf{Z}}_{\ell}^{\mathrm{T}}\right)\right) \tag{22}$$

其中, $\tilde{\mathbf{Z}}_{\ell}^{\mathrm{V}}$ 是多模态编码器输出的视觉表征, $\tilde{\mathbf{Z}}_{\ell}^{\mathrm{T}}$ 是多模态编码器输出的文本表征, 其中 y^{itm} 表示真值标签的二维独热表征.

2.4.3 图像文本对比学习损失

图文对比学习 (image-text contrastive learning, ITC) 旨在在融合之前学习更好的单模态表示. 它通过学习一种相似度函数, 使得成对图像和文字之间有更高的相似度分数. 其中, $\mathbf{Z}_{0}^{\mathrm{V}},\mathbf{Z}_{0}^{\mathrm{T}}$ 是单模态编码器得到的最后一层向量表征, $y^{\mathrm{v2t}}(\mathbf{Z}_{0}^{\mathrm{V}})$ 和 $y^{\mathrm{v2i}}(\mathbf{Z}_{0}^{\mathrm{T}})$ 表示真实的独热相似度, 负样本对的概率为 0, 正样本对的概率为 1. p^{v2t} 与 p^{v2i} 分别为图像

到文本和文本到图像的余弦相似度函数, 图文对比损失函数定义为 p 和 y 之间的交叉熵损失函数:

$$\mathcal{L}_{\text{itc}}=\frac{1}{2}\mathbb{E}_{(v,T)}\left[\mathrm{H}\left(y^{\text{v2t}}\left(\mathbf{Z}_0^{\text{V}}\right),p^{\text{v2t}}\left(\mathbf{Z}_0^{\text{V}}\right)\right)+\mathrm{H}\left(y^{\text{t2v}}\left(\mathbf{Z}_0^{\text{T}}\right),p^{\text{t2v}}\left(\mathbf{Z}_0^{\text{T}}\right)\right)\right] \tag{23}$$

MGFIN 最终的损失函数为:

$$\mathcal{L}=\mathcal{L}_{\text{vqa}}+\mathcal{L}_{\text{mlm}}+\mathcal{L}_{\text{itm}}+\mathcal{L}_{\text{itc}} \tag{24}$$

3 实验分析

3.1 数据集介绍

本文使用如下两种遥感 VQA 数据集来评估 MGFIN 模型的性能.

RSVQA-LR 是最早提出的低光谱分辨率遥感视觉问答数据集. LR 数据集包含 772 张 256×256 大小的图像, 来自分辨率为 10 m 的 Sentinel-2 卫星图像. 该数据集包含 77 232 个自然语言问题和相应的多选答案. 此数据集中分为 4 类问题, 即城市/农村、计数、存在和比较. 此外, 77.8%、11.1%、11.1% 的图像及其相关问题-答案对分别用于训练、验证和测试.

RSVQA-HR 是另一种最早的高光谱分辨率遥感视觉问答数据集, 采集自分辨率为 15 cm 的 USGS 航空 RGB 图像, 包括 10 659 张 512×512 大小的图像和 1 066 316 个问题-答案对. 对于 HR 数据集, 问题-答案样本分为 4 种类型: 面积、计数、存在和比较. 此外, 61.5%、11.2%、20.5%、6.8% 的图像样本及其相应的问题-答案对分别被分成训练集、验证集、测试集 1 和测试集 2. 其中测试集 1 覆盖了与训练集和验证集相似的区域, 而测试集 2 覆盖了其他区域的遥感信息, 且该区域在训练过程不可见.

3.2 参数设置

在视觉表征阶段, MGFIN 使用 Swin Transformer 提取多尺度视觉特征. 关于图像的输入大小, 本文采用原始图像大小, 即 RSVQA 数据集包含 256×256 的 LR 数据集和 512×512 的 HR 数据集. 尺度种类设置为 2, 关系类型设置为 1, 融合编码曾是设置为 2. 在语言表征阶段, MGFIN 采用 $\text{BERT}_{\text{base}}$ 作为文本编码器, 每个问题嵌入到 512 维向量中. 在训练阶段, 采用 Adam 优化器, 基础学习率分别为 1×10^{-4} (RSVQA-LR 数据集) 和 1×10^{-5} (HR 数据集). 训练和推理的批量大小在 RSVQA-LR 上设置为 70, 在 RSVQA-HR 数据集上设置为 16. 为了防止过拟合, 每个线性变换后都实现了 0.5 的 dropout. 分别在 50 个 epochs (LR 数据集)、30 个 epochs (HR 数据集) 后终止实验, 所有实验都是在一台配备 GeForce RTX 3090 显卡的服务器上运行. 为了全面评估 VQA 模型, 评估指标包括平均准确率 (average accuracy)、总体准确率 (overall accuracy) 和每种问题类型的准确率. 为了与以前的工作保持一致并显示系统偏差, 本文对比实验指标面向准确率的均值与标准差, 以体现实验结果指标的统计显著性, 继而忽略实验随机性扰动, 以明确模型效果.

3.3 对比实验结果及分析

在本节中, 本文将提出的 MGFIN 与当前列出的 4 种遥感 VQA 方法进行比较, 具体如下所示. 在 RSVQA-LR、RSVQA-HR 数据集上进行比较研究.

• RSVQA[23]: 一种基本方法, 简单地提取并组合视觉特征和问题特征以进行答案预测.

• EasyToHard[24]: 一种渐进式学习方法, 按照从易到难的顺序训练模型, 是一种契合模型感知与收敛的训练过程.

• Bi-modal[25]: 一种基于视觉语言 Transformer 的方法, 通过自我注意和协同注意机制对内部依赖性和跨模态依赖性进行建模.

• SHRNet[26]: 一种空间层次推理网络模型, 采用哈希空间位置编码和注意力机制引导推理来增强图像文本联合特征表示.

表 1 比较了以上模型在 RSVQA-LR 数据集上的性能. 展示了总体准确率、平均准确率和每种问题类型的准确率. 从比较结果来看, MGFIN 明显优于其他变体模型. 模型 MGFIN 实现了最佳的平均准确率 87.80%, 分别比新

的可用方 SHRNet 和原始基线 RSVQA 提高了 0.53% 和 7.74%. 对于不同的问题类别, MGFIN 在存在、比较和农村/城市上也表现最佳. 特别是对于农村/城市问题类型, MGFIN 模型相对于现有最佳模型实现了 3.19% 的准确率提高. 虽然本文的方法在计数和总体准确率类别上略逊于 SHRNet 方法, 但两种类型的准确率差距都不大. 由于 MGFIN 模型是针对空间多尺度的推理模型. 模型相对其他工作参数更多, 架构更复杂. 在小规模低分辨率数据集下, 对于多尺度特征的提取会受到分辨率的自然限制, 使得 MGFIN 抽取的多尺度特征不够鲁棒. 而在大规模高分数据集下, MGFIN 模型对多尺度特征的抽取, 对高阶特征推理融合的信息表征与挖掘优势才会有所体现.

表 1 RSVQA-LR 数据集上与现有先进方法的对比结果 (%)

类别	RSVQA[23]	EasyToHard[24]	Bi-modal[25]	SHRNet[26]	MGFIN (ours)
计数	67.01 (0.59)	69.22 (0.33)	72.22 (0.57)	**73.87 (0.22)**	71.25 (0.42)
存在	87.46 (0.06)	90.66 (0.24)	91.06 (0.17)	91.03 (0.13)	**91.64 (0.17)**
比较	81.50 (0.03)	87.49 (0.10)	91.16 (0.09)	90.48 (0.05)	**91.30 (0.04)**
农村/城市	90.00 (1.41)	91.67 (1.53)	92.66 (1.52)	94.00 (0.87)	**97.00 (1.39)**
平均准确度	81.49 (0.49)	84.76 (0.35)	86.78 (0.28)	87.34 (0.13)	**87.80 (0.24)**
总体准确度	79.08 (0.20)	83.09 (0.15)	85.56 (0.16)	**85.85 (0.28)**	85.56 (0.21)

表 2 和表 3 分别展示了 RSVQA-HR 数据集的两个测试集的比较结果. 在测试集 1 上, MGFIN 的总体准确率优于基线方法 RSVQA 2.68%, 并实现了最高的平均准确率 85.15%. MGFIN 在计数和存在类别上略逊于 SHRNet 方法, 但两种类型的准确率差距都不大. 在测试集 2 上, MGFIN 的总体准确率优于基线方法 RSVQA 4.15%, 平均准确率优于基线方法 RSVQA 4.22%. 然而 MGFIN 在计数和农村/城市类别上略逊于 SHRNet 方法, 但差距不大. 结合两表不难看出, 测试集 2 的总体性能不如测试集 1 的性能. 可能的原因是测试集 2 涵盖的区域与训练和验证集相对不同, 但在测试集 2 的性能也展示出 MGFIN 不俗的泛化性.

表 2 RSVQA-HR-Test01 数据集上与现有先进方法的对比结果 (%)

类别	RSVQA[23]	EasyToHard[24]	Bi-modal[25]	SHRNet[26]	MGFIN (ours)
计数	68.63 (0.11)	69.06 (0.13)	69.80 (0.09)	**70.04 (0.15)**	69.66 (0.09)
存在	90.43 (0.04)	91.39 (0.15)	92.03 (0.08)	**92.45 (0.11)**	92.38 (0.12)
比较	88.19 (0.08)	89.75 (0.12)	91.83 (0.00)	91.68 (0.09)	**92.02 (0.07)**
农村/城市	85.24 (0.05)	85.92 (0.19)	86.27 (0.05)	86.35 (0.13)	**86.54 (0.15)**
平均准确度	83.12 (0.03)	83.97 (0.06)	84.98 (0.05)	85.13 (0.08)	**85.15 (0.07)**
总体准确度	83.23 (0.02)	84.16 (0.05)	85.30 (0.05)	85.39 (0.05)	**85.46 (0.04)**

表 3 RSVQA-HR-Test02 数据集上与现有先进方法的对比结果 (%)

类别	RSVQA[23]	EasyToHard[24]	Bi-modal[25]	SHRNet[26]	MGFIN (ours)
计数	61.47 (0.08)	61.95 (0.08)	63.06 (0.11)	**63.42 (0.14)**	62.93 (0.10)
存在	86.26 (0.47)	87.97 (0.06)	89.37 (0.21)	89.81 (0.27)	**90.10 (0.22)**
比较	85.94 (0.12)	87.68 (0.23)	89.62 (0.29)	89.44 (0.23)	**89.88 (0.17)**
农村/城市	76.33 (0.50)	78.62 (0.23)	80.12 (0.39)	**80.37 (0.16)**	80.17 (0.33)
平均准确度	77.50 (0.29)	79.06 (0.15)	80.54 (0.16)	80.76 (0.21)	**80.77 (0.24)**
总体准确度	78.23 (0.25)	79.29 (0.15)	81.23 (0.15)	81.37 (0.19)	**81.48 (0.18)**

总之, 与其他方法相比, 所提出的模型 MGFIN 在以上 3 个数据集上都获得了最佳实验结果, 这证明了本文的模型在遥感 VQA 任务中的竞争优势. 从某种意义上说, 既有方法, 包括原始模型 RSVQA、改进版本 Bi-Modal 和最新的 SHRNet, 都与本文模型相关. 只因先前的方法和本文的方法都采用了类似的双通道结构和先进的技术, 包括 CNN 和 RNN 抑或多模态预训练模型 CLIP 来表征图像和文本. 然而, 此类的方法受到了视觉表示不足的限制, 并忽略了卫星图像中地理空间对象之间固有的细粒度视觉关系. 我们的模型带来显著改进的关键原因是 MGFIN

灵活地捕捉了多尺度对象特征之间的高阶关系, 并在语言线索的指导下进行空间分层推理. 另一方面, 虽然方法 Bi-modal 采用了先进的基于 Transformer 的编码器-解码器结构, 但可能不适合直接采用 CLIP 模型对遥感数据上的视觉和文本表示进行编码. 因为 CLIP 模型是在大规模自然图像-文本对通过预训练构建, 这与遥感图像-文本对存在明显的领域漂移. 也许正是由于这个原因, 本文提出的方法 MGFIN 比先进的基于 Transformer 的方法 Bi-modal 实现了更令人印象深刻的性能. 此外 SHRNet 通过哈希编码的方式引入空间信息, 可以很好地保留空间位置信息, 但其推理阶段还是采取了文本引导的注意力机制来得到融合表征, 最终得到的知识相对与 MGFIN 的关系建模稍显匮乏, 且缺乏多模态推理过程中的深度融合与各种模态知识的互相监督来为下游任务保驾护航.

3.4 消融实验结果及分析

本文的完整 VQA 模型的架构由多个基本模块组成, 也有重要的超参数需要讨论和验证.

3.4.1 模块消融实验

在本节中, 本文首先使用以下 MGFIN 的变体实施多个消融实验, 以验证每个组件对整体预测性能的贡献.

• MGFIN w/o 尺度: 该变体在视觉编码阶段删除多尺度表示模块, 使模型能够在属于单一尺度的对象外观特征之间执行关系推理.

• MGFIN w/o 位置编码: 该变体删除了空间位置嵌入, 这有助于保留视觉特征中的空间信息. 模型仅使用视觉外观特征在不同尺度上进行空间分层推理.

• MGFIN w/o 空间筛选: 该变体删除了多尺度空间筛选模块, 由于多尺度特征没有经过过滤, 将有助于保留视觉特征中的语义信息, 语义特征较过滤后更加完备, 但会保留冗余语义信息.

• MGFIN w/o 空间推理: 该变体删除了空间分层推理模块. 模型直接执行多模态特征融合, 而不考虑推理视觉空间关系.

• MGFIN w/o 桥融合: 该变体删除了 Bridge 交互模块, 该模块有助于增强视觉文本联合嵌入. 模型直接将视觉特征与问题特征组合以进行最终答案推理.

• MGFIN w/o 对齐监督: 该变体删除了对比学习、图文匹配和掩码语言模型这 3 个监督损失函数, 只保留视觉问答损失函数.

表 4 显示了 RSVQA-LR 数据集上 6 个变体的消融比较.

表 4 RSVQA-LR 数据集上的模块消融实验 (%)

模型变体	计数	存在	比较	农村/城市	平均准确度	总体准确度
MGFIN w/o 多尺度	66.78	88.80	88.71	92.00	84.07	82.31
MGFIN w/o 位置编码	**71.84**	91.03	91.07	95.00	87.24	85.44
MGFIN w/o 空间筛选	71.43	90.08	90.45	92.00	85.99	84.76
MGFIN w/o 空间推理	71.23	90.11	89.30	94.00	86.17	84.28
MGFIN w/o 桥融合	70.21	90.29	89.33	92.00	85.46	84.01
MGFIN w/o 对齐监督	70.58	89.98	89.28	94.00	85.96	84.02
MGFIN 完整模型	71.25	**91.64**	**91.30**	**97.00**	**87.80**	**85.56**

显然, 全模型 MGFIN 优于其他变体并取得了显著的改进. 与 MGFIN w/o 多尺度和 MGFIN w/o 位置编码相比, 全模型 MGFIN 具有更高的性能, 这表明多尺度视觉表示和特别设计的空间位置编码都是有效的. 此外, MGFIN w/o MS 的整体准确率下降更为显著, 这表明提取多尺度特征对于表示高分辨率图像具有更大的影响. 此外, 全模型 MGFIN 在 MGFIN w/o 空间推理、桥融合、对齐监督上实现了约 1%–2% 的改进, 证明采用以上 3 种模块对增强问题-图像联合嵌入也是有帮助的. 但在计数任务中, 位置编码与空间推理和空间筛选却起到负效果, 其共性在于增加模型对空间的感知, 增加模型复杂度, 间接证明计数任务对于模型复杂度和信息冗余度的敏感性. 且通过实验可证, 模型在缺少多尺度特征、空间位置编码、空间筛选的基础上缺乏对于空间信息的精准刻画, 在除计数的各项指标均逊于完整模型. 图 6 显示了 RSVQA-HR 数据集上 5 个变体模型对不同问题类型的准确率. 浅红色条形图反映了所提出的 MGFIN 的性能. 其余条形图对应于 MGFIN 的 4 个变体. 可以看出, MGFIN w/o 多尺

度 (黄色) 在所有 4 种问题类型上都表现最差, 这表明空间分层推理模块对遥感 VQA 有突出贡献. 与其余 4 个变体相比, 全模型 MGFIN 实现了最佳性能也证明了 MGFIN 中设计的组件的有效性.

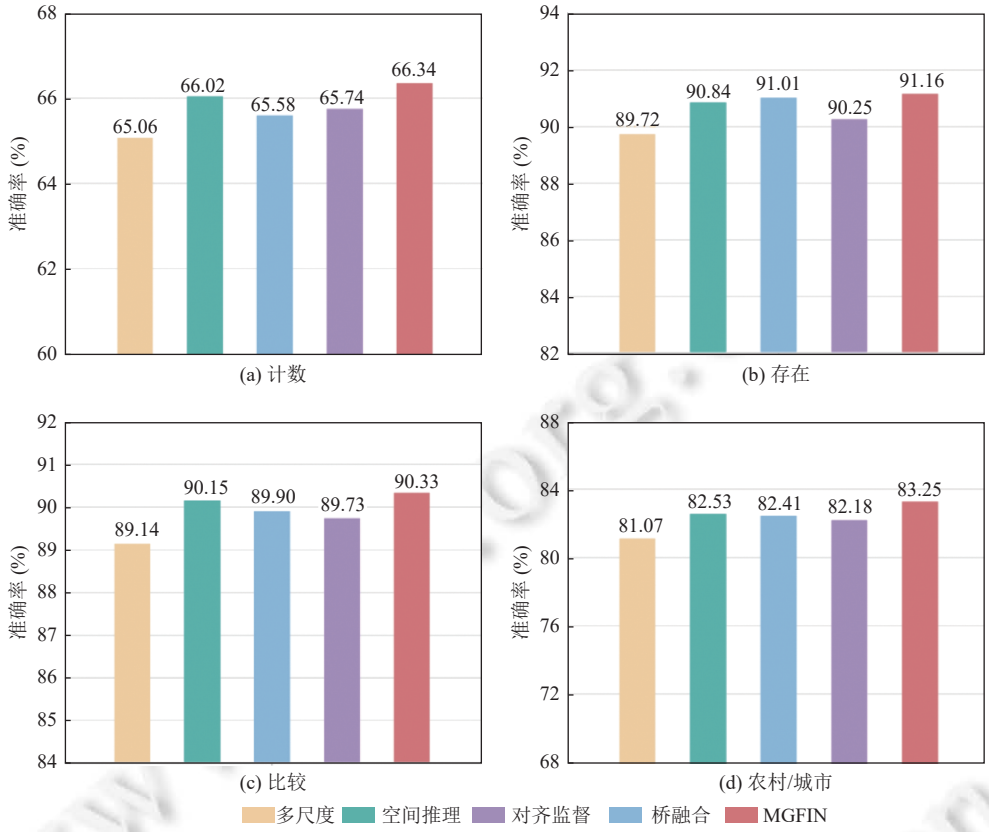


图 6 RSVQA-HR 数据集上 5 种变体模型对各类问题准确率

3.4.2 参数消融实验

如表 5 所示, 本文进行了多次实验以确定超参数的值, 结合公式 (11) 与公式 (12), 其中 S 与 M 分别指视觉特征的尺度种类数和多尺度推理中的高阶关系组合数. 除以上两组超参数, 本文还对融合编码器层数进行消融实验, 值得注意的是, 本文首先固定了最优参数, 然后改变其他参数并测试性能.

表 5 RSVQA-LR 数据集上的参数消融实验 (%)

参数设置	计数	存在	比较	农村/城市	平均准确度	总体准确度
尺度种类1	71.39	89.91	90.37	95.00	86.67	84.70
尺度种类2	71.25	**91.64**	**91.30**	**97.00**	**87.80**	85.56
尺度种类3	**72.81**	90.62	91.15	96.00	87.65	**85.65**
关系类型0	71.33	90.86	90.83	96.00	87.25	85.15
关系类型1	71.25	**91.64**	**91.30**	**97.00**	**87.80**	**85.65**
关系类型2	71.46	90.96	90.75	96.00	87.30	85.19
关系类型3	**72.44**	91.07	90.28	97.00	87.69	85.32
融合编码0	**71.43**	90.08	90.45	92.00	85.99	84.76
融合编码1	71.29	91.30	90.98	97.00	87.64	85.34
融合编码2	71.25	**91.64**	**91.30**	**97.00**	**87.80**	**85.56**
融合编码3	71.09	91.44	90.65	96.00	87.30	85.18

具体来说, 当 S (尺度种类) 的值过小时, 信息通过空间筛选模块的能力不足, 无法生成正确的答案. 同时, 恰当的 S 值有助于更好地描述更多的全局关系. 然而, 过大的 S 设置可能会在很大程度上削弱局部关系, 从而导致整体和单个项目性能较低. 换句话说, 大多数 VQA 问题都与局部关系的建模有关, 并且可以通过对局部关系的建模来回答, 而不是对全局关系的描述. 此外, 较大的 T 为该方法带来了更大的计算量. 因此, 本文通过参数消融把最终的 S 定义为 2, 以获得平衡的全局尺度与局部尺度表征.

在表 5 中, 还可以观察到高阶关系的推理会为问答系统提供知识增益, 然而, 更多的关系建模并不会提高性能. 归因于 VQA 任务的固有属性, 其中 RSVQA 通常涉及有限数量的对象进行回答. 因此, 更大的 M 并不是性能更好的必要条件, 驳杂的高阶关系建模会导致模型冗余以及软关系噪声的引入. 根据实验, 对于每个尺度, 推断所有可能的信息区域组合将带来巨大的计算负担. 因此, 本文定义 $M=1$ 以在性能和计算成本之间保持平衡.

对于编码器层数来说, 当模型不采用融合机制时, 模型在多数任务的效果都不理想, 但在计数任务上却取得不错的效果. 融合编码器可以促进单模态表征与多模态表征的融合, 但过多的层数会造成模型冗余, 过多的融合机制下驳杂的高阶关系建模会导致模型冗余以及软关系噪声的引入, 从而忽略候选区域的计数表征. 导致在映射过程中映射候选图像区域和问题的关系失真. 但根据其他任务的数据, 丰富的语义确实提升最终的映射效果. 因此, 本文定义融合编码层的层数为 2, 以在性能和模型参数之间保持平衡.

3.5 可视化展示

MGFIN 在空间层次推理融合阶段, 通过文本信息的引导注意力进一步增强了与问题相关的关系特征, 同时相应地抑制了不相关的关系特征. 为了发现包含在问题相关关系特征中并对视觉空间关系推理过程重要的图像区域, 本文通过累加分布在多个关系子集上的注意力权重并在可视化之前对其进行归一化来计算特征图上的注意力权重, 并对 3 个层次的注意力权重取平均值.

在图 7 中, 模型展示了在空间尺度推理阶段生成的注意力图的可视化.

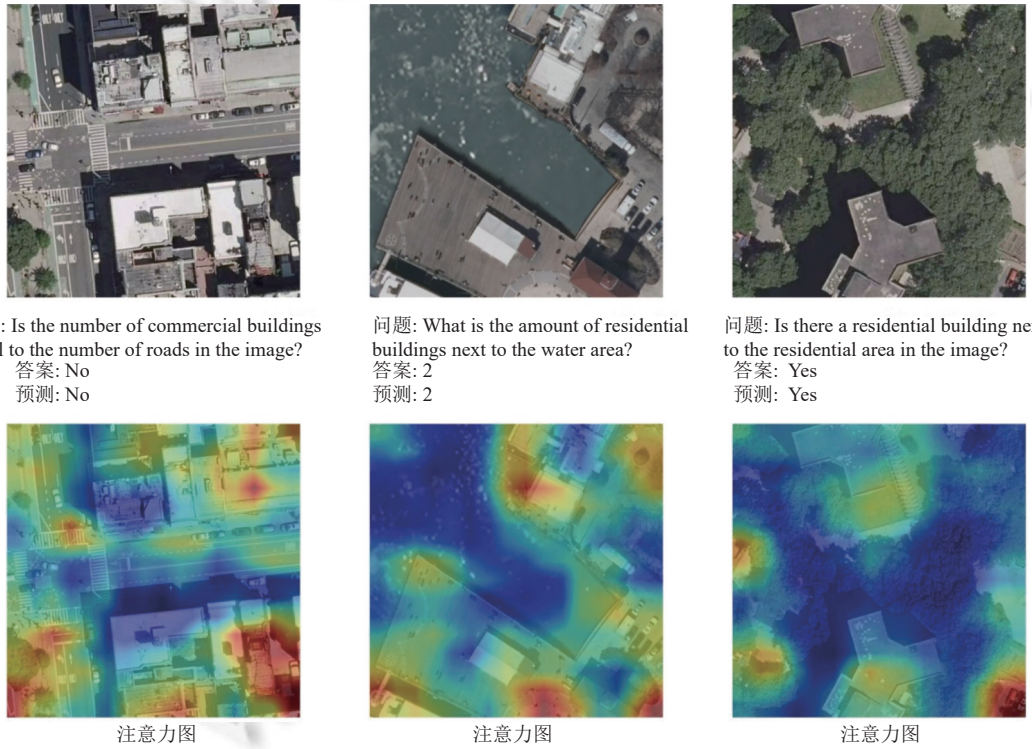


图 7 RSVQA-LR 数据集注意力图可视化

对于图 7 左边的示例, 问题涉及多个对象, 包括“住宅建筑”和“马路”, 需要比较它们的数量. 如注意力图所示, 网络关注所涉及的两个对象, 但具有不同程度的显著性, 然后推断出正确答案. 有趣的是, 在图 7 中间的示例中可以观察到类似的模式, 其中一类所涉及的对象“住宅建筑”用大面积突出显示, 而另一类所涉及的比较对象“水域”则用较小的权重和一小块区域集中. 此外, 对于图 7 右侧的查询“住宅建筑”的示例, 模型成功地突出了目标区域并过滤掉了不相关区域. 总之, 注意力图的可视化显示了文本问题和模型强调区域之间的一致性, 显著性程度显示了推断参考对象关系的证据.

4 总 结

本文提出了一种新颖的多尺度引导融合推理网络 (MGFIN), 为遥感视觉问答系统赋予了跨越多个尺度的视觉空间推理能力. 首先, 本文提出了一种基于空间位置和多尺度视觉表征模块, 用于编码嵌入空间位置信息的多尺度视觉特征. 其次, 通过多尺度空间层次推理模块学习文本引导下的多尺度视觉特征及其高阶语义关系, 得到丰富的语义表征. 最后, 本文通过引入多种监督信息, 结合交叉注意力机制获得融合充分的多模态表征继而推理出准确答案. 本文在两个公开可用的遥感 VQA 数据集上将提出的模型与既有方法进行比较, 并通过大量实验来评估 MGFIN 的有效性. 实验结果表明, MGFIN 在遥感 VQA 领域取得了最新的先进性能. 但受限于遥感域匮乏的知识表征与复杂的关系涌现, 在未来的工作中, 我们将探索对象级别注释的图像-问题-答案三元组, 或知识图谱等高阶知识驱动的工作, 以抽取目标的高阶特征, 继而更高效地问答推理. 此外, 由于遥感数据的稀缺性, 标注优质数据集或微调传统多模态大模型并将知识迁移至遥感域进行推理也是值得探索的方向.

References:

[1] Zador A, Escola S, Richards B, *et al*. Catalyzing next-generation Artificial Intelligence through NeuroAI. Nature Communications, 2023, 14(1): 1597. [doi: 10.1038/s41467-023-37180-x]

[2] Cavender-Bares J, Schneider FD, Santos MJ, Armstrong A, Carnaval A, Dahlin KM, Fatoyinbo L, Hurtt GC, Schimel D, Townsend PA, Ustin SL, Wang ZH, Wilson AM. Integrating remote sensing with ecology and evolution to advance biodiversity conservation. Nature Ecology & Evolution, 2022, 6(5): 506–519. [doi: 10.1038/s41559-022-01702-5]

[3] Zhang H, Li F, Liu SL, Zhang L, Su H, Zhu J, Ni LM, Shum HY. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In: Proc. of the 11th Int'l Conf. on Learning Representations. Kigali: OpenReview.net, 2023.

[4] Li MY, Cao CQ, Feng ZJ, Xu XK, Wu ZY, Ye SB, Yong JW. Remote sensing object detection based on strong feature extraction and prescreening network. IEEE Geoscience and Remote Sensing Letters, 2023, 20: 8000505. [doi: 10.1109/LGRS.2023.3236777]

[5] Li G, Li LL, Zhu H, Liu X, Jiao LC. Adaptive multiscale deep fusion residual network for remote sensing image classification. IEEE Trans. on Geoscience and Remote Sensing, 2019, 57(11): 8506–8521. [doi: 10.1109/TGRS.2019.2921342]

[6] Liu X, Jiao LC, Li LL, Cheng L, Liu F, Yang SY, Hou B. Deep multiview union learning network for multisource image classification. IEEE Trans. on Cybernetics, 2022, 52(6): 4534–4546. [doi: 10.1109/TCYB.2020.3029787]

[7] Liu X, Li LL, Liu F, Hou B, Yang SY, Jiao LC. GAFnet: Group attention fusion network for PAN and MS image high-resolution classification. IEEE Trans. on Cybernetics, 2022, 52(10): 10556–10569. [doi: 10.1109/TCYB.2021.3064571]

[8] Cheng G, Han JW, Lu XQ. Remote sensing image scene classification: Benchmark and state of the art. Proc. of the IEEE, 2017, 105(10): 1865–1883. [doi: 10.1109/JPROC.2017.2675998]

[9] Zhang F, Du B, Zhang LP. Saliency-guided unsupervised feature learning for scene classification. IEEE Trans. on Geoscience and Remote Sensing, 2015, 53(4): 2175–2184. [doi: 10.1109/TGRS.2014.2357078]

[10] Zhu H, Jiao LC, Ma WP, Liu F, Zhao W. A novel neural network for remote sensing image matching. IEEE Trans. on Neural Networks and Learning Systems, 2019, 30(9): 2853–2865. [doi: 10.1109/TNNLS.2018.2888757]

[11] Quan D, Wang S, Li Y, Yang BW, Huyan N, Chanussot J, Hou B, Jiao LC. Multi-relation attention network for image patch matching. IEEE Trans. on Image Processing, 2021, 30: 7127–7142. [doi: 10.1109/TIP.2021.3101414]

[12] Ma WP, Wen ZL, Wu Y, Jiao LC, Gong MG, Zheng YF, Liu L. Remote sensing image registration with modified SIFT and enhanced feature matching. IEEE Geoscience and Remote Sensing Letters, 2017, 14(1): 3–7. [doi: 10.1109/LGRS.2016.2600858]

[13] Ma AL, Wang JJ, Zhong YF, Zheng Z. FactSeg: Foreground activation-driven small object semantic segmentation in large-scale remote sensing imagery. IEEE Trans. on Geoscience and Remote Sensing, 2022, 60: 5606216. [doi: 10.1109/TGRS.2021.3097148]

[14] Zheng CY, Nie J, Wang ZX, Song N, Wang JY, Wei ZQ. High-order semantic decoupling network for remote sensing image semantic

segmentation. IEEE Trans. on Geoscience and Remote Sensing, 2023, 61: 5401415. [doi: 10.1109/TGRS.2023.3249230]

[15] Xie YX, Tian JJ, Zhu XX. Linking points with labels in 3D: A review of point cloud semantic segmentation. IEEE Geoscience and Remote Sensing Magazine, 2020, 8(4): 38–59. [doi: 10.1109/MGRS.2019.2937630]

[16] Li AJ, Jiao LC, Zhu H, Li LL, Liu F. Multitask semantic boundary awareness network for remote sensing image segmentation. IEEE Trans. on Geoscience and Remote Sensing, 2022, 60: 5400314. [doi: 10.1109/TGRS.2021.3050885]

[17] Zhang ZY, Zhang WK, Yan ML, Gao X, Fu K, Sun X. Global visual feature and linguistic state guided attention for remote sensing image captioning. IEEE Trans. on Geoscience and Remote Sensing, 2022, 60: 5615216. [doi: 10.1109/TGRS.2021.3132095]

[18] Zhao R, Shi ZW, Zou ZX. High-resolution remote sensing image captioning based on structured attention. IEEE Trans. on Geoscience and Remote Sensing, 2022, 60: 5603814. [doi: 10.1109/TGRS.2021.3070383]

[19] Li YP, Zhang XR, Gu J, Li C, Wang X, Tang X, Jiao LC. Recurrent attention and semantic gate for remote sensing image captioning. IEEE Trans. on Geoscience and Remote Sensing, 2022, 60: 5608816. [doi: 10.1109/TGRS.2021.3102590]

[20] Cheng QM, Zhou YZ, Fu P, Xu Y, Zhang L. A deep semantic alignment network for the cross-modal image-text retrieval in remote sensing. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14: 4284–4297. [doi: 10.1109/JSTARS.2021.3070872]

[21] Yuan ZQ, Zhang WK, Rong XE, Li X, Chen JL, Wang HQ, Fu K, Sun X. A lightweight multi-scale crossmodal text-image retrieval method in remote sensing. IEEE Trans. on Geoscience and Remote Sensing, 2022, 60: 5612819. [doi: 10.1109/TGRS.2021.3124252]

[22] Zheng G, Li XF, Zhou LZ, Yang JS, Ren L, Chen P, Zhang HG, Lou XL. Development of a gray-level co-occurrence matrix-based texture orientation estimation method and its application in sea surface wind direction retrieval from SAR imagery. IEEE Trans. on Geoscience and Remote Sensing, 2018, 56(9): 5244–5260. [doi: 10.1109/TGRS.2018.2812778]

[23] Lobry S, Marcos D, Murray J, Tuia D. RSVQA: Visual question answering for remote sensing data. IEEE Trans. on Geoscience and Remote Sensing, 2020, 58(12): 8555–8566. [doi: 10.1109/TGRS.2020.2988782]

[24] Yuan ZH, Mou LX, Wang Q, ZHU XX. From easy to hard: Learning language-guided curriculum for visual question answering on remote sensing data. IEEE Trans. on Geoscience and Remote Sensing, 2022, 60: 5623111. [doi: 10.1109/TGRS.2022.3173811]

[25] Bazi Y, Al Rahhal MM, Mekhalfi ML, Al Zuair MA, Melgani F. Bi-modal Transformer-based approach for visual question answering in remote sensing imagery. IEEE Trans. on Geoscience and Remote Sensing, 2022, 60: 4708011. [doi: 10.1109/TGRS.2022.3192460]

[26] Zhang ZX, Jiao LC, Li LL, Liu X, Chen PH, Liu F, Li YX, Guo ZC. A spatial hierarchical reasoning network for remote sensing visual question answering. IEEE Trans. on Geoscience and Remote Sensing, 2023, 61: 4400815. [doi: 10.1109/TGRS.2023.3237606]

[27] Antol S, Agrawal A, Lu JS, Mitchell M, Batra D, Zitnick CL, Parikh D. VQA: Visual question answering. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision. Santiago: IEEE, 2015. 2425–2433. [doi: 10.1109/ICCV.2015.279]

[28] Goyal Y, Khot T, Summers-Stay D, Batra D, Parikh D. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 6325–6334. [doi: 10.1109/CVPR.2017.670]

[29] Fukui A, Park DH, Yang D, Rohrbach A, Darrell T, Rohrbach M. Multimodal compact bilinear pooling for visual question answering and visual grounding. In: Proc. of the 2016 Conf. on Empirical Methods in Natural Language Processing. Austin: Association for Computational Linguistics, 2016. 457–468. [doi: 10.18653/v1/D16-1044]

[30] Kim JH, On KW, Lim W, Jeonghee Kim, Ha JW, Zhang BT. Hadamard product for low-rank bilinear pooling. In: Proc. of the 5th Int'l Conf. on Learning Representations. Toulon: OpenReview.net, 2017.

[31] Yu Z, Yu J, Fan JP, Tao DC. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 1839–1848.

[32] Yang ZC, He XD, Gao JF, Deng L, Smola A. Stacked attention networks for image question answering. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 21–29. [doi: 10.1109/CVPR.2016.10]

[33] Anderson P, He XD, Buehler C, Teney D, Johnson M, Gould S, Zhang L. Bottom-up and top-down attention for image captioning and visual question answering. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 6077–6086. [doi: 10.1109/CVPR.2018.00636]

[34] Ren SQ, He KM, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137–1149. [doi: 10.1109/TPAMI.2016.2577031]

[35] Song JK, Zeng PP, Gao LL, Shen HT. From pixels to objects: Cubic visual attention for visual question answering. In: Proc. of the 27th Int'l Joint Conf. on Artificial Intelligence. Stockholm: IJCAI.org, 2018. 906–912. [doi: 10.24963/ijcai.2018/126]

[36] Yu Z, Yu J, Cui YH, Tao DC, Tian Q. Deep modular co-attention networks for visual question answering. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 6274–6283. [doi: 10.1109/CVPR.2019.00644]

[37] Chappuis C, Zermatten V, Lobry S, Le Saux B, Tuia D. Prompt-RSVQA: Prompting visual context to a language model for Remote Sensing Visual Question Answering. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops. New Orleans: IEEE, 2022. 1371–1380. [doi: 10.1109/CVPRW56347.2022.00143]

[38] Yuan ZH, Mou LC, Xiong ZT, Zhu XX. Change detection meets visual question answering. IEEE Trans. on Geoscience and Remote Sensing, 2022, 60: 5630613. [doi: 10.1109/TGRS.2022.3203314]

[39] Santoro A, Raposo D, Barrett DGT, Malinowski M, Pascanu R, Battaglia P, Lillicrap T. A simple neural network module for relational reasoning. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 4974–4983. [doi: 10.5555/3295222.3295250]

[40] Zhou BL, Andonian A, Oliva A, Torralba A. Temporal relational reasoning in videos. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 831–846. [doi: 10.1007/978-3-030-01246-5_49]

[41] Le TM, Le V, Venkatesh S, Tran T. Hierarchical conditional relation networks for video question answering. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 9969–9978. [doi: 10.1109/CVPR42600.2020.00999]

[42] Hu H, Gu JY, Zhang Z, Dai JF, Wei YC. Relation networks for object detection. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 3588–3597. [doi: 10.1109/CVPR.2018.00378]

[43] Mou LC, Hua YS, Zhu XX. A relation-augmented fully convolutional network for semantic segmentation in aerial scenes. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 12408–12417. [doi: 10.1109/CVPR.2019.01270]

[44] Liu Z, Hu H, Lin YT, Yao ZL, Xie ZD, Wei YX, Ning J, Cao Y, Zhang Z, Dong L, Wei FR, Guo BN. Swin Transformer V2: Scaling up capacity and resolution. In: Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 11999–12009. [doi: 10.1109/CVPR52688.2022.01170]

[45] Xu X, Wu CF, Rosenman S, Lal V, Che WX, Duan N. BridgeTower: Building bridges between encoders in vision-language representation learning. In: Proc. of the 37th AAAI Conf. on Artificial Intelligence. Washington: AAAI, 2023. 10637–10647. [doi: 10.1609/aaai.v37i9.26263]

[46] Li JM, Selvaraju RR, Gotmare AD, Joty S, Xiong CM, Hoi SCH. Align before fuse: Vision and language representation learning with momentum distillation. In: Proc. of the 35th Conf. on Neural Information Processing Systems. NeurIPS, 2021. 9694–9705.

赵恩源(1995－), 男, 博士生, 主要研究领域为深度学习, 多模态智能系统.

王鑫(1981－), 男, 博士, 助理研究员, CCF 高级会员, 主要研究领域为媒体大数据分析, 机器学习, 多媒体智能.

宋宁(1996－), 男, 博士生, 主要研究领域为科学人工智能, 海洋多模态智能计算.

郑程予(1994－), 女, 博士生, 主要研究领域为人工智能, 跨模态大数据分析.

聂婕(1984－), 女, 博士, 副教授, CCF 专业会员, 主要研究领域为多模态智能计算, 人工智能, 海洋环境预测预报.

魏志强(1969－), 男, 博士, 教授, CCF 高级会员, 主要研究领域为海洋大数据智能挖掘, 高性能计算.