



工学硕士学位论文

基于多尺度旋转不变与知识图谱优化的零样本遥感 图像语义分割

Zero-shot Semantic Segmentation of Remote Sensing Images
Based on Multi-scale Rotation Invariance and Knowledge Graph
Optimization

作 者 吕东瑞

导 师 崔宾阁 教授

山东科技大学

二〇二五年五月

中图分类号 TP751

学校代码 10424

UDC 004

密 级 公开

山东科技大学

工学硕士学位论文

基于多尺度旋转不变与知识图谱优化的零样本遥感图像语义分割

Zero-shot Semantic Segmentation of Remote Sensing Images

Based on Multi-scale Rotation Invariance and Knowledge Graph

Optimization

作 者 吕东瑞 入学时间 2022 年 9 月

导 师 崔宾阁 职 称 教授

申请学位 工学硕士 所在学院 计算机科学与工程学院

学科（类别） 计算机应用技术 方向（领域） 人工智能

答辩日期 2025 年 5 月 18 日 答辩委员会主席

答辩委员会成员

提交日期

学位论文使用授权声明

本人完全了解山东科技大学有关保留、使用学位论文的规定，同意本人所撰写的学位论文的使用授权按照学校的管理规定处理。

作为申请学位的条件之一，学校有权保留学位论文并向国家有关部门或其指定机构送交论文的电子版和纸质版；有权将学位论文的全部或部分内容编入有关数据库发表，并可以以电子、网络及其他数字媒体形式公开出版；允许学校档案馆和图书馆保留学位论文的纸质版和电子版，可以使用影印、缩印或扫描等复制手段保存和汇编学位论文；为教学和科研目的，学校档案馆和图书馆可以将公开的学位论文作为资料在档案馆、图书馆等场所或在校园网上供校内师生阅读、浏览。

（保密的学位论文在解密后适用本授权）

作者签名：吕东瑞

日 期： 2025 年 月 日

导师签名：崔宾阁

日 期： 2025 年 月 日

学位论文原创性声明

本人呈交给山东科技大学的学位论文，除所列参考文献和世所公认的文献外，全部是本人攻读学位期间在导师指导下的研究成果。除文中已经标明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对本文的研究做出贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。若有不实之处，本人愿意承担相关法律责任。

学位论文作者签名：吕东瑞

2025 年 月 日

学位论文审查认定书

研究生吕东瑞在规定的学习年限内，按照培养方案及个人培养计划，完成了课程学习，成绩合格，修满规定学分；在我的指导下完成本学位论文，论文中的观点、数据、表述和结构为我所认同，论文撰写格式符合学校的相关规定，同意将本论文作为申请学位论文。

导师签名：崔宾阁
日期：

摘要

遥感图像语义分割技术在城市规划、环境监测、灾害评估及农业管理等领域展现出重要应用价值。然而，遥感图像因其高空间分辨率、多尺度目标分布、多视角成像及复杂场景特性，对传统监督学习方法构成了显著挑战。特别是在新场景或新类别出现时，获取标注数据的高昂成本极大限制了模型的适应性。为克服这一难题，零样本语义分割技术应运而生，旨在无需标注数据支持的情况下实现不可见类别的精准分割。本文基于 CLIP 模型的跨模态语义对齐能力，结合遥感图像的独特特性，提出了一系列创新方法以提升零样本语义分割性能，主要研究内容涵盖以下两个方面：

(1) 针对遥感图像中目标尺度差异大和多视角成像导致的旋转不变性不足问题，本文提出了一种多尺度增强和旋转不变性网络 **MSRINet**，用于实现遥感图像零样本语义分割。**MSRINet** 通过三大核心模块提升模型性能：首先，全局-局部信息增强模块结合 CLIP 的全局语义信息与 ResNet 的局部细节信息，利用特征金字塔网络实现全局-局部特征融合，构建包含丰富多尺度信息的特征表示，有效应对目标尺度差异；其次，旋转不变性特征优化模块通过多角度旋转生成等变特征，结合空间注意力机制提取旋转不变特征，提升模型对多视角目标的鲁棒性；最后，类别关系建模模块利用自注意力机制挖掘类别间的复杂语义关系，增强模型对复杂遥感场景的理解能力。

(2) 针对遥感图像中地物类别语义的领域特异性问题，本文提出了一种基于知识图谱和信息瓶颈理论的零样本语义分割方法 **KGIBNet**，以增强模型对不可见类别的语义区分能力和视觉特征表达的精准性。**KGIBNet** 通过两大核心模块实现性能提升：首先，属性知识增强模块利用知识图谱的属性信息，通过提示模板策略将类别三元组转化为自然语言句子，并结合 CLIP 文本编码器生成文本语义嵌入，显著提升对领域特异性类别的语义表达能力；其次，基于信息瓶颈的知识融合模块通过图卷积网络和多头自注意力机制分别对空间关系和属性信息进行建模，并利用信息瓶颈理论将文本语义信息和视觉信息进行融合，筛选与多模态融合的相关信息，滤除模态间的冗余信息，从而增强特征表示的精准性。

在 GID5、OpenEarthMap 和 LoveDA 数据集上的实验结果表明，**MSRINet** 提升了对不同尺度目标的表征能力以及对多视角目标的鲁棒性，**KGIBNet** 则有效缓解了地物类别语义领域特异性难题。两者分别从视觉特征优化和语义知识增强两个维度出发，共同应对遥感图像零样本语义分割的核心挑战，为复杂地物分割任务提供了鲁棒、可解释的解决方案，展现了其在实际应用中的潜力。

关键词：遥感图像；零样本语义分割；旋转不变性；知识图谱；信息瓶颈理论

Abstract

Remote sensing image semantic segmentation technology has demonstrated significant application value in fields such as urban planning, environmental monitoring, disaster assessment, and agricultural management. However, due to the high spatial resolution, multi-scale target distribution, multi-view imaging, and complex scene characteristics of remote sensing images, it poses significant challenges to traditional supervised learning methods. Especially when new scenes or new categories emerge, the high cost of obtaining annotated data greatly limits the adaptability of the model. To overcome this challenge, zero-shot semantic segmentation technology has emerged, aiming to achieve accurate segmentation of unseen categories without the support of annotated data. Based on the cross-modal semantic alignment ability of the CLIP model and combined with the unique characteristics of remote sensing images, this thesis proposes a series of innovative methods to improve the performance of zero-shot semantic segmentation. The main research content covers the following two aspects:

(1) Aiming at the problems of large target scale differences in remote sensing images and insufficient rotational invariance caused by multi-view imaging, this thesis proposes a multi-scale enhancement and rotational invariance network, MSRINet, for zero-shot semantic segmentation of remote sensing images. MSRINet improves the model performance through three core modules: Firstly, the Global-Local Information Enhancement Module combines the global semantic information of CLIP with the local detail information of ResNet, and uses the feature pyramid network to achieve cross-scale feature fusion, constructing a feature representation containing rich multi-scale information, effectively dealing with target scale differences; Secondly, the rotational invariance feature optimization module generates equivariant features through multi-angle rotation, and combines the spatial attention mechanism to extract rotational invariant features, improving the model's robustness to multi-view targets; Finally, the category relationship modeling module uses the self-attention mechanism to mine the complex semantic relationships between categories, enhancing the model's understanding ability of complex remote sensing scenes.

(2) Aiming at the problem of the domain specificity of the semantic of land object categories in remote sensing images, this thesis proposes a zero-shot semantic segmentation method based on knowledge graph and information bottleneck theory, KGIBNet, to enhance the model's semantic discrimination ability for unseen categories and the accuracy of visual feature expression. KGIBNet achieves performance improvement through two core modules: Firstly, the Attribute Knowledge Enhancement Module uses the structured information of the knowledge graph, transforms category triples into natural language sentences through the prompt template strategy, and combines the CLIP text encoder to generate text semantic embeddings, significantly

improving the semantic expression ability for domain-specific categories; Secondly, the knowledge fusion module based on the information bottleneck uses the graph convolutional network and multi-head self-attention mechanism to model the spatial relationships and attribute information respectively, and uses the information bottleneck theory to fuse the text semantic information and visual information, screen the information relevant to multi-modal fusion, and filter out the redundant information between modalities, thus enhancing the accuracy of feature representation.

The experimental results on the GID5, OpenEarthMap, and LoveDA datasets show that MSRINet improves the representation ability for targets of different scales and the robustness to multi-view targets, while KGIBNet effectively alleviates the problem of the domain specificity of the semantic of land object categories. Starting from the two dimensions of visual feature optimization and semantic knowledge enhancement respectively, the two jointly address the core challenges of zero-shot semantic segmentation of remote sensing images, providing a robust and interpretable solution for complex land object segmentation tasks and demonstrating their potential in practical applications.

Keywords: Remote sensing image; Zero-shot segmentation; Rotational invariance; Knowledge graph; Information bottleneck theory

目 录

图清单	I
表清单	III
变量注释表	IV
1 绪 论	1
1.1 研究背景与意义 (Background and significance of research)	1
1.2 国内外研究现状 (Research status at home and abroad)	2
1.3 主要研究内容 (Main research contents)	4
1.4 论文章节安排 (Arrangement of thesis chapters)	5
2 相关技术	7
2.1 卷积神经网络 (Convolutional neural networks)	7
2.2 特征提取主干网络 (The backbone network of feature extraction)	9
2.3 对比语言-图像预训练 CLIP (Contrastive Language-Image Pretraining)	11
2.4 遥感地学知识图谱 (Remote Sensing Geoscience Knowledge Graph)	12
2.5 信息瓶颈理论 (Information Bottleneck Theory)	14
2.6 零样本语义分割 (Zero-Shot Semantic Segmentation)	14
2.7 本章小结 (Summary)	16
3 多尺度增强和旋转不变性零样本语义分割方法	17
3.1 引言 (Foreword)	17
3.2 多尺度增强和旋转不变性网络 MSRINet (Multi - scale Enhancement and Rotational Invariance Network)	18
3.3 实验结果与分析 (Experimental results and analysis)	22
3.4 本章小结 (Summary)	34
4 基于知识图谱和信息瓶颈理论的零样本语义分割方法	36
4.1 引言 (Foreword)	36
4.2 知识图谱与信息瓶颈网络 KGIBNet (Knowledge Graph and Information Bottleneck Network)	37
4.3 实验结果与分析 (Experimental results and analysis)	42
4.4 本章小结 (Summary)	51
5 总结与展望	52
5.1 总结 (Conclusion)	52

5.2 展望 (Prospect)	52
参考文献	54
作者简历	59
致 谢	60
学位论文数据集	61

Contents

List of Figures	I
List of Tables	III
List of Variables	IV
1 Introduction	1
1.1 Background and significance of research	1
1.2 Research status at home and abroad	2
1.3 Main research contents	4
1.4 Arrangement of thesis chapters	5
2 Related technology	7
2.1 Convolutional neural networks	7
2.2 The backbone network of feature extraction	9
2.3 Contrastive Language-Image Pretraining	11
2.4 Remote Sensing Geoscience Knowledge Graph	12
2.5 Information Bottleneck Theory	14
2.6 Zero - Shot Semantic Segmentation	14
2.7 Summary	16
3 Multi-scale Enhancement and Rotation Invariance-based Zero-shot Segmentation Method	17
3.1 Foreword	17
3.2 Multi - scale Enhancement and Rotational Invariance Network	18
3.3 Experimental results and analysis	22
3.4 Summary	34
4 Knowledge Graph and Information Bottleneck-based Zero-Shot Semantic Segmentation	36
4.1 Foreword	36
4.2 Knowledge Graph and Information Bottleneck Network	37
4.3 Experimental results and analysis	42
4.4 Summary	51
5 Conclusion and prospect	52
5.1 Conclusion	52
5.2 Prospect	52
References	54

Author's Resume	59
Acknowledgements	60
Thesis Data Collection	61

图清单

图序号	图名称	页码
图 2.1	二维卷积过程示意图	6
Fig.2.1	Schematic diagram of the two-dimensional convolution process	6
图 2.2	常见激活函数	7
Fig.2.2	Common activation functions	7
图 2.3	最大池化与平均池化示意图	8
Fig.2.3	The schematic diagram of max pooling and average pooling	8
图 2.4	残差连接示意图	9
Fig.2.4	Schematic diagram of residual connection	9
图 2.5	FPN 网络结构	9
Fig.2.5	The architecture of the Feature Pyramid Network (FPN)	9
图 2.6	Vision Transformer 网络结构	10
Fig.2.6	The architecture of the Vision Transformer network	10
图 2.7	CLIP 模型的训练过程	11
Fig.2.7	Training process of the CLIP model	11
图 2.8	使用 CLIP 模型进行零样本推理	11
Fig.2.8	Zero-shot inference using the CLIP model	11
图 2.9	图卷积网络	13
Fig.2.9	Graph Convolutional Network	13
图 2.10	零样本语义分割	14
Fig.2.10	Zero-Shot Semantic Segmentation	14
图 3.1	MSRINet 整体架构	17
Fig.3.1	The overall architecture of MSRINet	17
图 3.2	全局-局部信息增强模块	18
Fig.3.2	Global-Local Information Enhancement Module	18
图 3.3	旋转不变性特征优化模块	20
Fig.3.3	Rotation Invariance Feature Optimization Module	20
图 3.4	类别关系建模模块	21
Fig.3.4	Category Relationship Modeling Module	21
图 3.5	零样本语义分割数据集制作流程	25
Fig.3.5	The process of making a zero-shot semantic segmentation dataset	25
图 3.6	GID5 数据集上 MSRINet 与不同方法的预测结果	27
Fig.3.6	Prediction results of MSRINet and different methods on the GID5 dataset	27
图 3.7	OpenEarthMap 数据集上 MSRINet 与不同方法的预测结果	28
Fig.3.7	Prediction results of MSRINet and different methods on the OpenEarthMap dataset	28
图 3.8	LoveDA 数据集上 MSRINet 与不同方法的预测结果	29
Fig.3.8	Prediction results of MSRINet and different methods on the LoveDA dataset	29

图 3.9	模型中间过程特征图可视化	31
Fig.3.9	Visualization of the Feature Maps in the Intermediate Process of the Model	31
图 4.1	仅使用 CLIP 通用类别标签预测的原图像和热力图	35
Fig.4.1	The original image and the heatmap predicted using only the CLIP general category labels	35
图 4.2	KGIBNet 整体架构	37
Fig.4.2	The overall architecture of KGIBNet	37
图 4.3	属性知识增强模块	38
Fig.4.3	Attribute Knowledge Enhancement Module	38
图 4.4	基于信息瓶颈的知识融合	39
Fig.4.4	Knowledge Fusion Based on Information Bottleneck	39
图 4.5	GID5 数据集上 KGIBNet 与不同方法的预测结果	42
Fig.4.5	Prediction results of KGIBNet and different methods on the GID5 dataset	42
图 4.6	OpenEarthMap 数据集上 KGIBNet 与不同方法的预测结果	43
Fig.4.6	Prediction results of KGIBNet and different methods on the OpenEarthMap dataset	43
图 4.7	LoveDA 数据集上 KGIBNet 与不同方法的预测结果	45
Fig.4.7	Prediction results of KGIBNet and different methods on the LoveDA dataset	45
图 4.8	原始类别文本嵌入和知识增强嵌入的可视化比较	47
Fig.4.8	Visual Comparison of the Original Categorical Text Embeddings and the Knowledge-Enhanced Embeddings	47
图 4.9	不同数据集的超参数分析	49
Fig.4.9	Hyperparameter Analysis of Different Datasets	49
图 4.10	GID5 数据集不同方法的 mIoU 和推理速度对比	50
Fig.4.10	Comparison of mIoU and Inference Speed of Different Methods on the GID5 Dataset	50

表清单

表序号	表名称	页码
表 3.1	实验遥感数据集介绍	24
Table 3.1	Introduction to Experimental Remote Sensing Datasets	24
表 3.2	零样本遥感数据集统计表格	25
Table 3.2	Statistical Table of Zero-shot Remote Sensing Datasets	25
表 3.3	GID5 数据集上 MSRINet 与不同方法的精度比较	28
Table 3.3	Comparison of the accuracies between MSRINet and different methods on the GID5 dataset	28
表 3.4	OpenEarthMap 数据集上 MSRINet 与不同方法的精度比较	29
Table 3.4	Comparison of the accuracies between MSRINet and different methods on the OpenEarthMap dataset	29
表 3.5	LoveDA 数据集上 MSRINet 与不同方法的精度比较	30
Table 3.5	Comparison of the accuracies between MSRINet and different methods on the LoveDA dataset	30
表 3.6	MSRINet 的消融实验	31
Table 3.6	The abalition experiment of MSRINet	31
表 3.7	CLIP 中间层选择对不同数据集分割性能的影响	33
Table 3.7	Impact of Selected CLIP Intermediate Layers on the Segmentation Performance of Different Datasets	33
表 3.8	RIFOM 中旋转度数对不同数据集分割性能的影响	33
Table 3.8	The Influence of the Rotation Degrees in RIFOM on the Segmentation Performance of Different Datasets	33
表 4.1	GID5 数据集上 KGIBNet 与不同方法的精度比较	43
Table 4.1	Comparison of the accuracies between KGIBNet and different methods on the GID5 dataset	43
表 4.2	OpenEarthMap 数据集上 KGIBNet 与不同方法的精度比较	44
Table 4.2	Comparison of the accuracies between KGIBNet and different methods on the OpenEarthMap dataset	44
表 4.3	LoveDA 数据集上 KGIBNet 与不同方法的精度比较	45
Table 4.3	Comparison of the accuracies between KGIBNet and different methods on the LoveDA dataset	45
表 4.4	KGIBNet 的消融实验	46
Table 4.4	The abalition experiment of KGIBNet	46
表 4.5	模型复杂度与性能对比	49
Table 4.5	Model Complexity and Performance Comparison	49

变量注释表

变量	注释	初现页
\tilde{A}	带自环的邻接矩阵	12
C^s	可见类别	14
C^u	不可见类别	14
d	嵌入维度	18
d_k	键向量维度	10
d_r	降维后嵌入维度	21
\tilde{D}	度矩阵	12
E^V	CLIP 提取的视觉嵌入	17
E^L	CLIP 提取的文本嵌入	19
E_{GL}	全局-局部增强嵌入	19
E_{GL}^V	全局-局部增强视觉嵌入	19
E_{kg}	属性知识嵌入	38
E_{kg}^L	文本语义嵌入	38
E_{triple_i}	三元组嵌入	38
F_{CRM}	CRMM 输出特征图	21
F_{clip}^1	CLIP 第 1 层中间层的视觉特征图	18
F_{clip}^3	CLIP 第 3 层中间层的视觉特征图	18
F_{clip}^5	CLIP 第 5 层中间层的视觉特征图	18
\mathcal{F}_{clip}	CLIP 提取的多尺度特征集合	18
F_{cos}	语义匹配特征图	19
F_{COS}	语义匹配特征体	19
\mathcal{F}_{GL}	全局-局部融合特征	19
F_{IB}	类别知识语义特征图	40
F_{inv}	旋转不变特征图	20
F_{KGIB}	知识瓶颈优化特征图	36
\mathcal{F}_{res}	ResNet 提取的多尺度特征集合	19
F_{res}^{C1}	ResNet C1 中间层的视觉特征图	19
F_{res}^{C2}	ResNet C3 中间层的视觉特征图	19
F_{res}^{C3}	ResNet C2 中间层的视觉特征图	19
F_{RIFO}	RIFOM 输出特征图	20
F_{RIFO_s}	轻量级 F_{RIFO} 特征图	21
F_s	空间增强特征图	20
F_{θ_i}	旋转等变特征图	20
F_w	加权类别特征	21
H	特征图的高度	17
$H^{(l)}$	第 l 层节点特征矩阵	12
h_i	头实体	37

I	输入的遥感图像	17
N_t	三元组数量	37
K	键向量	10
n	序列长度	10
N_c	类别数	19
Q	查询向量	10
r_i	关系实体	37
T	压缩表示	13
T_c	知识图谱三元组	37
t_i	尾实体	37
V	值向量	10
W	特征图的宽度	17
w_c	自学习的类别权重	38
w_s	空间权重描述符	20
$W^{(l)}$	可学习的权重矩阵	12
W_K	键矩阵	10
W_Q	查询矩阵	10
W_V	值矩阵	10
X	输入向量	10
Y	输出目标	13
z_{att}	优化后的属性嵌入	40
z_{spa}	优化后的空间嵌入	40
α	自学习的加权权重	19
β	权衡超参数	13
θ_i	旋转角度	20
Θ	旋转角度集合	20
σ	非线性激活函数	12
$\Phi^L(\cdot)$	CLIP 文本编码器	19
$\Phi^V(\cdot)$	CLIP 视觉编码器	18

1 绪论

1 Introduction

1.1 研究背景与意义 (Background and significance of research)

随着遥感技术的快速发展，遥感图像在环境监测、城市规划、灾害评估、农业管理等领域发挥着越来越重要的作用^[1]。遥感图像具有高空间分辨率、覆盖范围广、数据量大等特点，能够提供地球表面多维、多时相的信息，成为地理空间分析和决策支持的重要数据来源^[2]。然而，其复杂的背景、多样的地物类型以及动态变化的场景对图像分割技术提出了更高的要求^[3]。例如，高分辨率遥感图像中地物类别的边界模糊、尺度差异显著，而低分辨率图像则可能因光谱混杂导致类别区分困难^[4]。传统的图像分割方法通常依赖于大量标注数据进行监督学习，然而在遥感领域，数据标注不仅耗时耗力，还面临着新场景和新类别的泛化能力不足的问题。特别是在面对不可见过的新类别（即训练数据中未包含的类别）时，传统方法往往表现不佳，这限制了遥感图像分割技术在实际应用中的灵活性。例如，在灾害监测中，新型灾害区域可能涉及未标注的地物类别，而传统模型难以快速适应这种动态需求^[5]。

近年来，零样本学习作为一种新兴的机器学习范式，为解决上述问题提供了新的思路^[6]。零样本学习通过利用语义信息（如类别描述、属性或词嵌入）在训练阶段建立类别间的知识迁移机制，从而实现对不可见过类别的识别与分割^[7]。其核心思想是通过构建视觉特征与语义信息之间的映射关系，使得模型能够在训练阶段学习的知识迁移到不可见类别上^[8]。在遥感图像领域，零样本语义分割的研究逐渐兴起^[9]，其核心目标是在不依赖新类别标注数据的情况下，利用已有的知识对遥感图像中的未知地物类别进行精确分割。这一技术对于提升遥感图像分析的自动化水平、降低数据标注成本以及增强模型在新场景下的适应能力具有重要意义^[10]。例如，在生态保护领域，零样本语义分割技术可用于快速识别未标注的新型植被或入侵物种；在城市规划中，可用于自动提取新开发的建筑物或道路网络。

尽管零样本语义分割在自然图像领域已取得了一定进展^{[11][12][13]}，但在遥感图像领域的应用仍面临诸多挑战。例如，遥感图像中地物的光谱特性复杂、多尺度特征显著，且不同类别间的语义关系较为复杂，这些都对零样本语义分割的模型设计和性能优化提出了更高的要求。具体而言，遥感图像中的地物类别往往具有高度的异质性，例如同一类别的地物在不同地域或不同成像条件下可能表现出显著的光谱差异，而不同类别的地物在某些场景下可能具有相似的视觉特征，这增加了类别间语义关系的建模难度。此外，遥感图像的零样本语义分割研究尚处于起步阶段，现有方法在泛化能力、分割精度以及实际应用场景的适配性方面仍存在较大提升空间。例如，当前的零样本语义分割模型在

处理遥感图像中的小目标分割或复杂背景下的类别区分时，往往难以达到理想的效果。因此，针对遥感图像零样本语义分割问题的深入研究，不仅具有重要的学术价值，还对推动遥感技术的智能化发展具有深远的现实意义。通过探索遥感图像零样本语义分割的关键技术，可以为遥感领域的智能化分析提供新的解决方案，并为相关实际应用场景提供技术支撑。

1.2 国内外研究现状（Research status at home and abroad）

随着遥感技术的飞速发展和计算机视觉领域的突破，遥感图像的解译方法经历了从传统人工解译到自动化、智能化解译的深刻变革。近年来，基于机器学习和深度学习的遥感图像解译技术在效率和精度上取得了显著提升，尤其是在语义分割、零样本学习以及开放词汇分割等方向上，国内外研究者提出了众多创新性方法。

（1）遥感图像语义分割

遥感图像语义分割是遥感解译领域的核心任务之一，其目标是对图像中的每个像素进行类别标注，以实现地物类别的精确提取。近年来，深度学习技术的引入显著提升了语义分割的性能，尤其是在高分辨率遥感图像的处理中，研究者们提出了多种创新性框架和模型。

Diakogiannis 等人提出了一种用于遥感数据语义分割的深度学习框架 ResUNet-a，通过结合 UNet 编码器/解码器、残差连接、空洞卷积和多任务推理，达到了当时的最优性能^[14]。崔宾阁等人提出了一种结合深度学习和植被指数的滨海湿地信息提取网络 MFVNet，通过增强多尺度特征提取模块和语义特征融合，在黄河口滨海湿地高分二号遥感影像上实现了优于传统机器学习和深度语义分割网络的信息提取性能^[15]。Li 等人提出了一种用于高分辨率遥感图像语义分割的多注意力网络 MANet，通过引入线性复杂度的核注意力机制和通道注意力模块，增强了特征图的长距离依赖建模能力，并在两个大规模遥感数据集上展示了优于现有方法的性能^[16]。Zhao 等人提出了一种基于注意力机制的遥感图像语义分割网络 SSAtNet，通过设计金字塔注意力池化模块和池化索引校正模块优化多尺度特征利用和细节信息恢复，在 ISPRS Vaihingen 数据集上达到了当时的最优性能^[17]。Ma 等人提出了一种用于遥感语义分割的多级多模态融合变换器 FTransUNet，通过集成卷积神经网络和视觉变换器，结合自适应互增强注意力层和自注意力层^[18]。

（2）零样本学习

尽管语义分割技术在已知类别上的表现优异，但其对训练数据中不可见类别的泛化能力仍存在不足，而零样本学习作为一种解决不可见类别识别问题的范式，为遥感图像解译提供了新的思路。零样本学习通过构建视觉特征与语义信息之间的映射关系，实现对不可见类别的知识迁移，近年来在遥感领域的应用逐渐增多。

Li 等人提出了一种基于对比视觉-语言监督的零样本遥感场景分类方法（RS-CLIP），

通过引入伪标签技术和课程学习策略，在多个遥感基准数据集上实现了零样本和少样本场景分类的性能提升^[19]。Sumbul 等人研究了遥感图像中的细粒度对象识别和零样本学习问题，提出了一种通过卷积神经网络提取图像特征并结合语义信息实现不可见类别识别的兼容性函数方法，并在高分辨率航空数据中验证了其有效性^[20]。Li 等人提出了一种针对高空间分辨率遥感图像的零样本场景分类（ZSSC）方法，通过构建语义导向图和标签传播算法实现从可见类到不可见类的知识迁移，并在分类结果中利用稀疏学习进行标签精炼，显著提升了零样本分类性能^[21]。Chaudhuri 等人提出了一种标签缺失零样本学习（LDZSL）训练协议，用于遥感图像的跨传感器数据检索，通过设计双层 Siamese 网络实现多光谱和合成孔径雷达（SAR）图像的跨模态检索，在少量标签数据条件下展现了优于现有零样本学习框架的性能^[22]。

（3）零样本语义分割和开放词汇语义分割

在零样本学习的基础上，零样本语义分割和开放词汇语义分割进一步扩展了遥感图像解译的应用场景，旨在实现对不可见类别的像素级分割或开放类别集的语义分割。这些技术结合了视觉-语言模型、注意力机制和多尺度特征提取方法，为遥感图像的自动化分析提供了更高的灵活性和泛化能力。

Huang 等人提出了一种用于零样本遥感实例分割的框架 ZoRI，通过设计判别增强分类器、知识保持适配策略和先验注入预测模块，解决了遥感图像中类间相似性和类内差异大的问题^[23]。Bane 等人提出了一种基于注意力机制的零样本遥感无监督分割方法 RS-A2M，通过利用视觉变换器和零样本检测器，优化了分割效率与精度^[24]。Li 等人提出了一种无需训练的遥感图像开放词汇语义分割方法 Segearth-ov，通过设计简单高效的上采样器 SimFeatUp 和消除全局偏差的减法操作，增强了特征的空间信息恢复能力^[25]。Cao 等人提出了一种专为遥感图像设计的开放词汇语义分割框架，通过引入旋转聚合相似性计算模块和多尺度特征上采样，解决了图像方向变化和尺度差异问题^[26]。Huang 等人提出了一种用于无人机航空图像开放词汇理解的视觉-语言框架，通过结合 Mixup 技术、GPT 文本描述生成和 U-Net-Mamba 架构，提升了对不可见类别的识别与分割能力^[27]。

（4）知识推理与数据学习的结合

在遥感图像解译的诸多方法中，单纯依赖数据驱动的深度学习模型往往难以处理复杂场景下的地物类别区分问题，而知识推理与数据学习的有机结合逐渐成为解决这一问题的重要趋势。通过引入先验知识，研究者们能够在提高解译精度的同时增强模型的可解释性，为遥感大数据的智能化处理提供了新的视角。

Li 等人提出了一种地理知识驱动的遥感图像表征学习方法 GeoKR，通过将全球土地覆盖产品和地理位置作为地理知识，为表征学习和网络预训练提供监督信息，并设计高效的预训练框架消除成像时间和分辨率差异带来的监督噪声^[28]。Li 等人提出了一种结合知识图谱与深度学习的新型遥感图像解译范式，通过构建遥感知识图谱 RS-KG 并应用于

零样本场景分类、可解释语义分割和大尺度场景图生成任务，提升了解译模型的精度、泛化能力和可解释性^[29]。Gun 等人提出了一种结合知识图谱和最优分割算法的高分辨率遥感图像分类方案，通过利用知识图谱保留分割对象间的空间关系并进行推理，解决光谱相似对象的误分类问题^[30]。Cui 等人提出了一种基于知识和地理对象的图卷积网络 KGCCN，通过将超像素块视为图网络节点并结合先验知识与空间相关性进行信息聚合，克服了样本上下文信息失真的问题^[31]。Wang 等人提出了一种多模态知识图谱感知的深度图注意力网络 MMKDGA，通过构建遥感图像的多模态知识图谱并结合深度关系注意力机制，实现对图像多种属性和视觉信息的整合，提升了主动推荐的精度^[32]。

1.3 主要研究内容（Main research contents）

本文针对遥感图像零样本语义分割中的核心挑战——多尺度目标分布、旋转不变性不足以及地物类别的领域特异性语义问题，基于 CLIP 模型的跨模态语义对齐能力，提出了一系列创新方法以提升分割性能。研究聚焦于解决传统监督学习方法在标注数据稀缺场景下的局限性，旨在实现无需标注数据支持的不可见类别分割。提出了多尺度增强和旋转不变性网络（MSRINet）以及基于知识图谱和信息瓶颈理论网络（KGIBNet）两种创新方法。

针对遥感图像中目标尺度差异大和多视角成像导致的旋转不变性不足问题，提出了一种多尺度增强和旋转不变性网络（MSRINet）。该方法通过全局-局部信息增强模块（GLIEM）融合 CLIP 的全局语义特征与 ResNet 的局部细节特征，利用特征金字塔网络实现全局-局部特征整合；通过旋转不变性特征优化模块（RIFOM）生成等变特征并提取旋转不变特征，提升模型对多视角目标的适应性；通过类别关系建模模块（CRMM）挖掘类别间的语义关系，增强复杂场景的理解能力。MSRINet 旨在从视觉特征层面优化零样本分割性能，为后续语义增强奠定基础。

针对遥感图像中地物类别的领域特异性语义及视觉特征复杂性问题，提出了一种基于知识图谱和信息瓶颈理论的零样本语义分割方法（KGIBNet）。该方法通过属性知识增强模块（AKEM）利用知识图谱的结构化信息生成领域特异性语义嵌入；通过基于信息瓶颈的知识融合模块（KFBIB）结合图卷积网络和多头自注意力机制优化空间关系和属性信息，并运用信息瓶颈理论筛选关键特征，滤除冗余噪声。KGIBNet 在 MSRINet 的基础上进一步提升了不可见类别的语义区分能力和特征表达精度，形成视觉特征优化与语义增强的协同解决方案。

此外，研究在 GID5、OpenEarthMap 和 LoveDA 等数据集上验证了两种方法的性能，与 ZegFormer、zsseg、SAN、SED 和 CAT-Seg 等方法进行对比分析，并通过消融实验评估各模块的有效性，验证方法的鲁棒性与泛化能力。本文旨在为遥感图像零样本语义分割提供高效、鲁棒的解决方案，降低标注成本，提升模型在新场景下的适应性，为城市规

划、环境监测等领域提供技术支持。

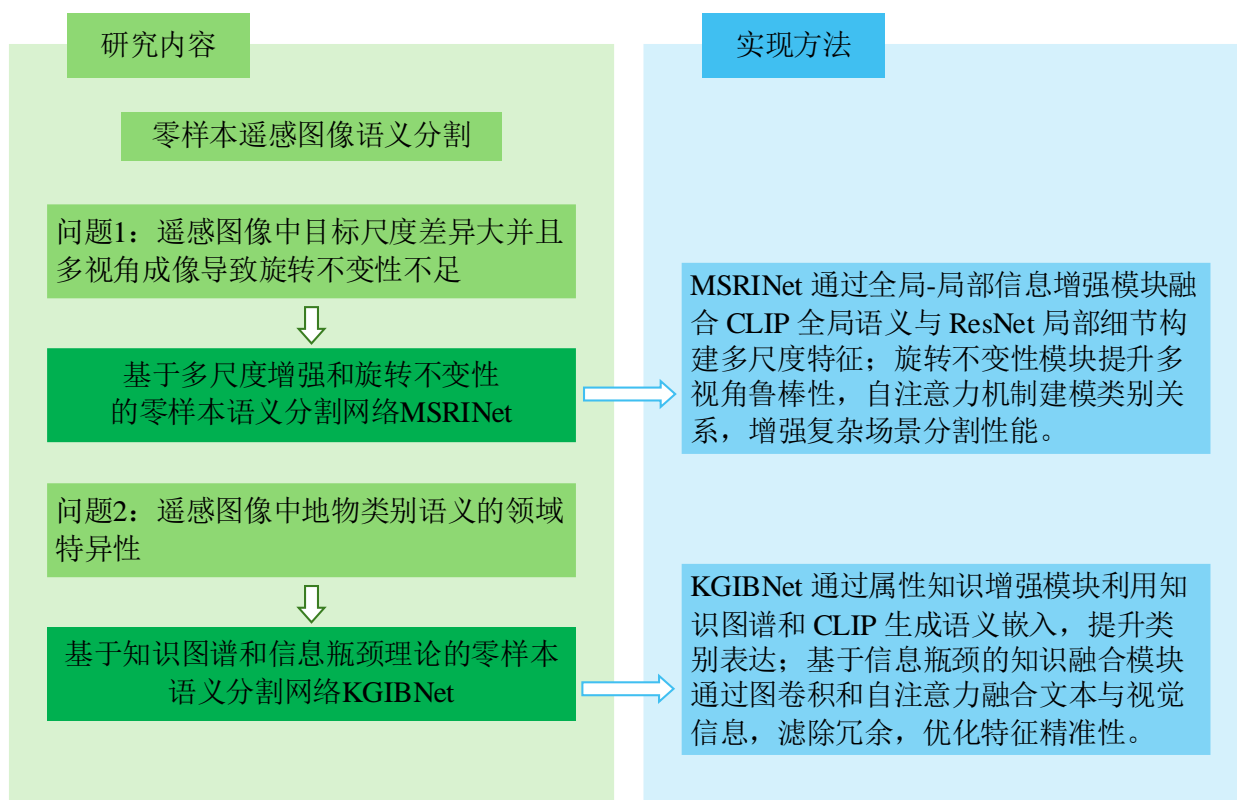


图 1.1 整体研究方案图

Fig.1.1 The diagram of general research scheme

1.4 论文章节安排 (Arrangement of thesis chapters)

本文的研究内容设计为五章，分别如下：

第一章为绪论，阐述遥感图像语义分割的研究背景与意义，分析传统监督学习方法的局限性及零样本语义分割的应用价值，综述国内外相关研究现状，指出当前挑战与空白，并提出本文的主要研究内容与创新点，概述后续章节安排。

第二章介绍相关技术，详细讲解卷积神经网络（CNN）的基本原理及特征提取主干网络（如 ResNet、FPN、ViT）的应用，分析 CLIP 模型的跨模态预训练机制及其在零样本任务中的优势，探讨遥感地学知识图谱的结构特性及图卷积网络（GCN）在知识推理中的作用，介绍信息瓶颈理论及其相关应用，并梳理零样本语义分割的技术基础与研究现状，为方法设计奠定理论基础。

第三章详细描述 MSRINet 的设计与实现过程，包括全局-局部信息增强模块（GLIEM）、旋转不变性特征优化模块（RIFOM）和类别关系建模模块（CRMM）的原理与功能。结合实验结果与消融分析，验证 MSRINet 在解决多尺度目标分布和旋转不变性问题上的有效性。

第四章针对地物类别语义的领域特异性问题，提出基于知识图谱和信息瓶颈理论的

KGIBNet 方法，阐述其整体架构及属性知识增强模块（AKEM）与基于信息瓶颈的知识融合模块（KFBIB）的实现细节，通过实验对比与消融分析评估其在不可见类别分割中的性能优势与鲁棒性。

第五章为总结与展望，归纳本文的研究成果与 MSRINet、KGIBNet 的创新点及其应用价值，分析方法的不足，提出未来研究方向，包括轻量化设计、知识图谱自动化构建及多模态融合等，以进一步提升技术性能与实用性。

2 相关技术

2 Related technology

2.1 卷积神经网络 (Convolutional neural networks)

卷积神经网络 (Convolutional Neural Network, 简称 CNN) 是一种专门用于处理具有网格结构数据的深度学习模型, 特别在计算机视觉任务中表现出色, 例如图像分类、目标检测和图像分割等^[33]。它的设计灵感来源于人类视觉系统, 通过多层网络结构自动提取数据的特征, 而无需手动设计。CNN 的核心思想在于通过局部连接、权值共享和池化操作减少参数数量, 提高计算效率, 同时保留数据的空间结构信息。典型的 CNN 由输入层、卷积层、激活层、池化层、全连接层和输出层等多个组成部分构成, 每一层在特征提取和处理中扮演不同角色, 最终实现从低级特征 (如边缘、纹理) 到高级特征 (如物体形状、语义信息) 的逐步抽象。

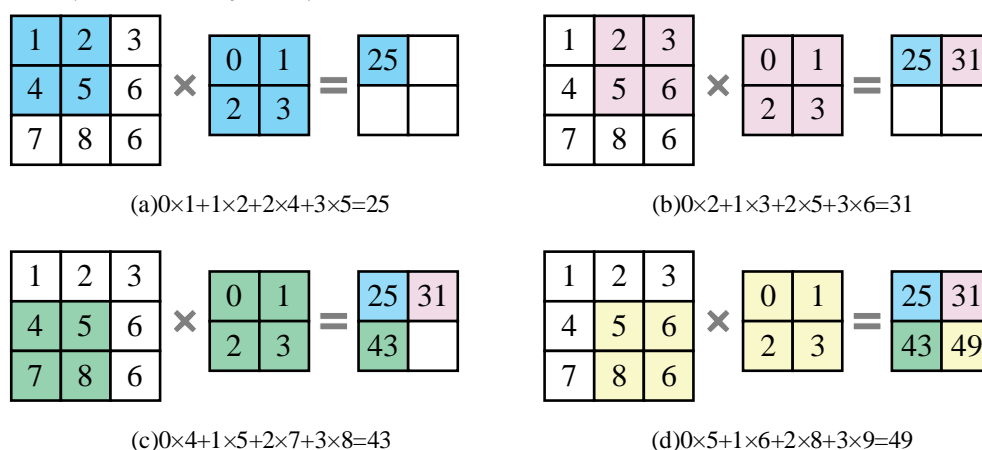


图 2.1 二维卷积过程示意图

Fig.2.1 Schematic diagram of the two-dimensional convolution process

卷积层是 CNN 的核心部分, 通过卷积核 (一个小的权重矩阵, 例如 3×3 、 5×5 或 7×7) 在输入数据上滑动, 计算局部区域的加权和, 生成特征图^[34]。卷积核本质上是一个特征提取器, 其权重在训练过程中通过反向传播算法自动学习, 从而能够捕捉输入数据中的特定模式, 例如边缘、角点或纹理等。卷积操作具有局部连接和权值共享的特点: 局部连接意味着每个卷积核只关注输入数据的局部区域, 而不是整个输入, 这大大减少了参数数量和计算量; 权值共享则意味着同一个卷积核的权重在整个输入数据上共享, 即无论卷积核滑动到哪个位置, 其权重保持不变, 这不仅进一步减少了参数数量, 还使得卷积操作对输入数据的位置变换具有一定的鲁棒性。卷积层的输出特征图数量由卷积核的数量决定, 每个卷积核生成一张特征图, 捕捉不同的特征模式^[35]。

激活层紧随卷积层之后, 通过引入非线性激活函数增强网络的学习能力, 使其能够拟合复杂的非线性模式^[36]。如果没有激活层, 卷积操作仅仅是线性变换, 多层堆叠的卷

积层仍然等价于一个线性模型，无法解决复杂的模式识别问题。激活函数的作用是对卷积层的输出进行非线性映射，常见的激活函数包括 ReLU (Rectified Linear Unit)、Sigmoid、Tanh 和 Leaky ReLU 等。其中，ReLU 因其简单性和加速收敛的特性而被广泛使用。ReLU 通过将负值置为 0，保留正值，不仅引入了稀疏性（即部分神经元输出为 0），减少了计算量，还在一定程度上缓解了梯度消失问题，从而加速了网络的训练过程。

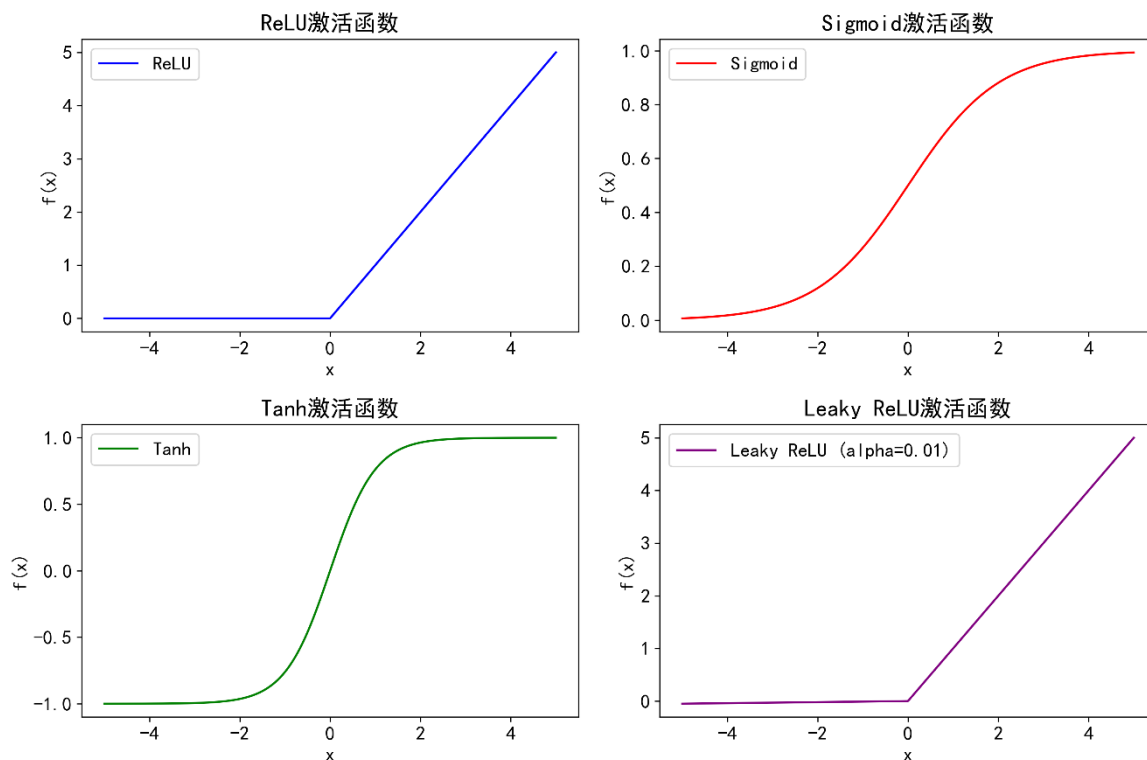


图 2.2 常见激活函数

Fig.2.2 Common activation functions

池化层在 CNN 中用于对特征图进行下采样，通过缩小空间维度（即宽度和高度）来降低计算复杂度，同时增强模型对平移、旋转、缩放等变换的鲁棒性。池化操作的核心思想是通过在特征图上滑动一个窗口，对窗口内的值进行汇总，从而在保留主要特征的同时减少数据量^[37]。常见的池化方式包括最大池化 (Max Pooling) 和平均池化 (Average Pooling)。最大池化通过取窗口内的最大值，突出显著特征（如边缘或角点），这在特征提取中非常有效，因为它倾向于保留最强的信号；平均池化则通过取窗口内的平均值，平滑特征图，减少噪声的影响，但可能会丢失一些细节信息^[38]，示意图如图 2.3。

全连接层负责将前几层提取的高级特征整合起来，用于最终的分类、回归或其他任务。全连接层的工作原理是将卷积层或池化层的输出特征图展平 (Flatten) 为一个一维向量，然后通过权重矩阵和偏置进行线性变换，生成一个新的向量。全连接层的每个神经元与输入的所有神经元相连，类似于传统神经网络的多层感知机 (MLP)，因此其参数数量通常较多^[39]。

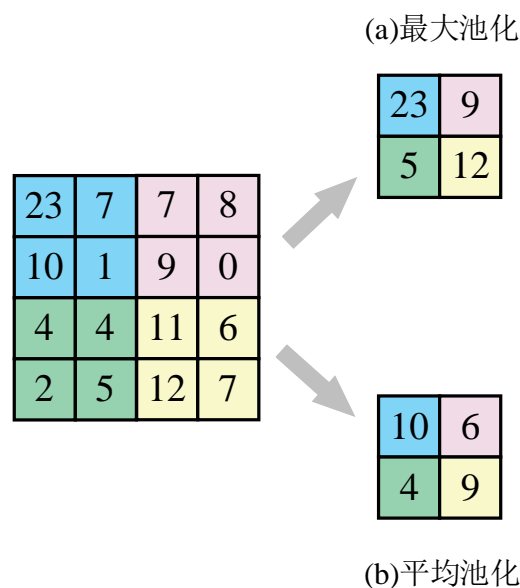


图 2.3 最大池化与平均池化示意图

Fig.2.3 The schematic diagram of max pooling and average pooling

2.2 特征提取主干网络（The backbone network of feature extraction）

在计算机视觉领域，主干网络（Backbone Network）作为特征提取的核心组件，扮演着至关重要的角色。它们通过对输入图像进行多层次的特征提取，为后续任务（如图像分类、目标检测、语义分割等）提供丰富的语义信息和空间信息。随着深度学习的发展，卷积神经网络（CNN）长期占据主干网络的主导地位，其中 ResNet 以其残差连接的设计成为经典代表^[40]；而为了应对多尺度目标的挑战，特征金字塔网络（FPN）作为一种增强模块被广泛应用。与此同时，近年来基于 Transformer 的架构逐渐兴起^[41]，Vision Transformer（ViT）凭借其全局建模能力，开辟了视觉特征提取的新方向。

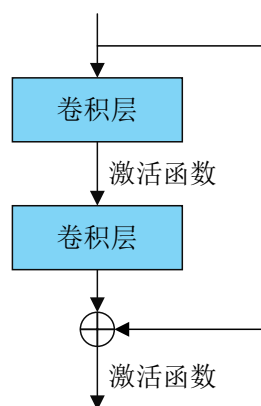


图 2.4 残差连接示意图

Fig.2.4 Schematic diagram of residual connection

ResNet（Residual Network，残差网络）是一种经典的深度卷积神经网络架构，旨在解决深层网络中的梯度消失问题，从而使网络可以更深层次地堆叠。它通过引入“残差连

接”（Residual Connection）来实现这一目标，即在网络的每一层中，不仅学习直接的特征映射，还学习输入和输出之间的残差。这种设计使得网络更容易优化，并且在实践中表现出了强大的特征提取能力。ResNet 有多种变体，例如 ResNet-18、ResNet-50、ResNet-101 等，数字表示网络的层数。ResNet 在图像分类、目标检测、语义分割等任务中被广泛用作主干网络，尤其是在需要提取深层视觉特征时表现出色。

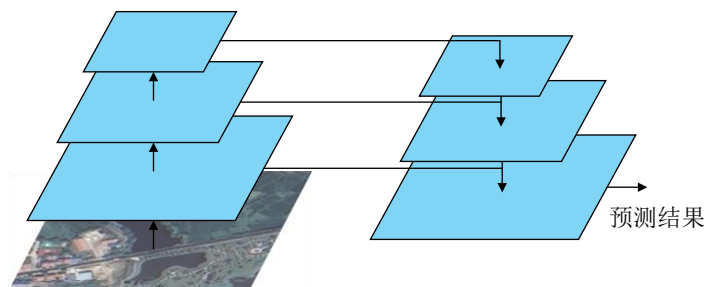


图 2.5 FPN 网络结构

Fig.2.5 The architecture of the Feature Pyramid Network (FPN)

FPN（Feature Pyramid Network，特征金字塔 2 塔网络）是一种用于多尺度特征提取的网络结构，通常与主干网络（如 ResNet）结合使用，以增强对不同尺度目标的检测能力^[42]。FPN 的核心思想是通过构建一个自顶向下的特征金字塔结构，将深层网络的高语义信息与浅层网络的高分辨率信息相结合。具体来说，FPN 在主干网络的多个层级上提取特征，并通过横向连接和上采样操作，将高层特征与低层特征融合，从而生成一组具有强语义和多尺度分辨率的特征图。这种设计特别适用于目标检测和实例分割等任务，能够有效提升对小目标的检测性能。FPN 并不是一个独立的主干网络，而是对主干网络特征提取能力的增强模块。

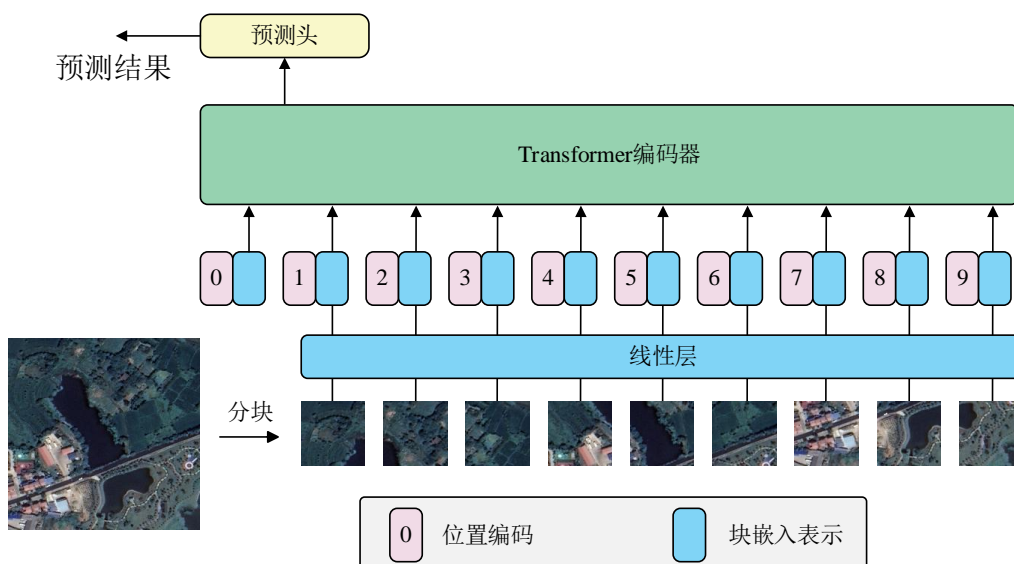


图 2.6 Vision Transformer 网络结构

Fig.2.6 The architecture of the Vision Transformer network

ViT（Vision Transformer）是一种基于 Transformer 架构的视觉特征提取网络，最初由 Google 提出，旨在将自然语言处理领域的 Transformer 模型迁移到计算机视觉任务中^[43]。

与传统的卷积神经网络（如 ResNet）不同，ViT 摒弃了卷积操作，而是将输入图像分割成固定大小的图像块（Patch），并将这些图像块展平后作为序列输入到 Transformer 中。通过自注意力机制（Self-Attention），ViT 能够捕捉图像中全局范围内的依赖关系，从而提取全局特征。ViT 在训练时通常需要大量数据（如 ImageNet-21k）进行预训练，以克服局部感受野的局限性。ViT 的优点在于其对长距离依赖关系的建模能力，尤其在高分辨率图像任务中表现出色，已逐渐成为视觉主干网络的一个重要方向，广泛应用于分类、检测和分割等任务。

自注意力机制（Self-Attention Mechanism）是 Transformer 模型的核心，用于捕捉输入序列中各元素间的全局依赖关系^[44]。其基本原理是将输入向量 X 通过三个可学习矩阵 W_Q 、 W_K 、 W_V 映射为查询 $Q = XW_Q$ 、键 $K = XW_K$ 、值 $V = XW_V$ 。注意力分数通过点积计算并归一化，公式为：

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (2.1)$$

其中 d_k 为键向量的维度，用于缩放以稳定梯度。多头自注意力通过在多个子空间并行计算并拼接结果，进一步增强建模能力。自注意力机制的优势在于全局感受野，但计算复杂度为 $O(n^2)$ ， n 为序列长度，因此在视觉任务中常通过局部窗口等优化降低成本。

2.3 对比语言-图像预训练 CLIP（Contrastive language-image pretraining）

CLIP（Contrastive Language-Image Pretraining，对比语言-图像预训练）是 OpenAI 于 2021 年开发的一种多模态深度学习模型，旨在通过联合训练图像编码器和文本编码器，将图像和文本映射到同一嵌入空间，从而捕捉两者之间的语义关系。其核心创新在于使用对比学习（Contrastive Learning），通过最大化匹配的图像-文本对的余弦相似度，同时最小化不匹配对的相似度，实现高效的多模态表示学习^[45]。

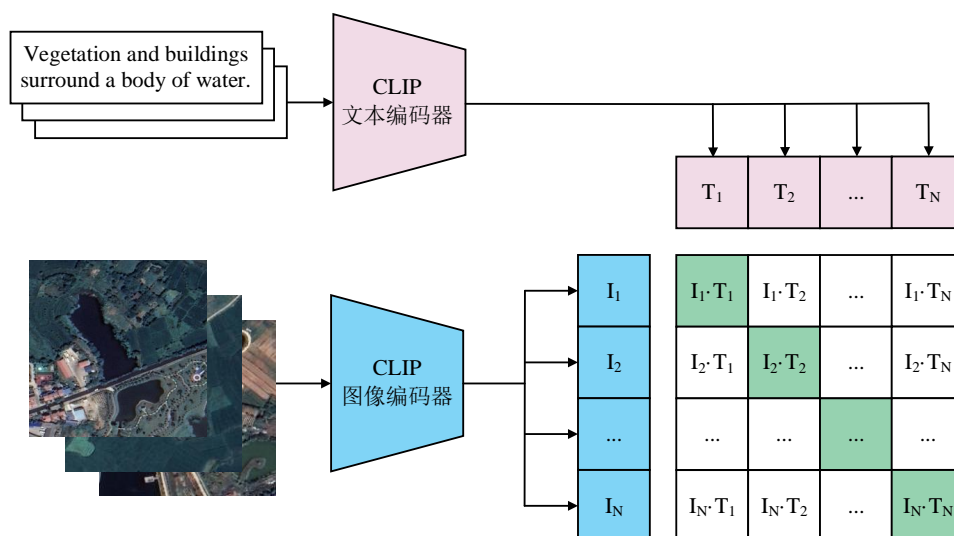


图 2.7 CLIP 模型的训练过程

Fig.2.7 Training process of the CLIP model

CLIP 在训练时使用了 4 亿对图像-文本数据，数据来源于互联网，训练成本高昂，例如在 AWS 上训练需约 100 万美元的计算资源。CLIP 的架构包括图像编码器（可采用 ResNet 或 Vision Transformer）和文本编码器（基于 Transformer），两者输出的嵌入向量在共享空间中对齐。CLIP 的一个显著特点是零样本学习（Zero-Shot Learning），即无需针对特定任务进行额外训练，就能通过文本描述对图像进行分类或检索，广泛应用于图像分类、内容检索、文本生成图像等多种场景^[46]。

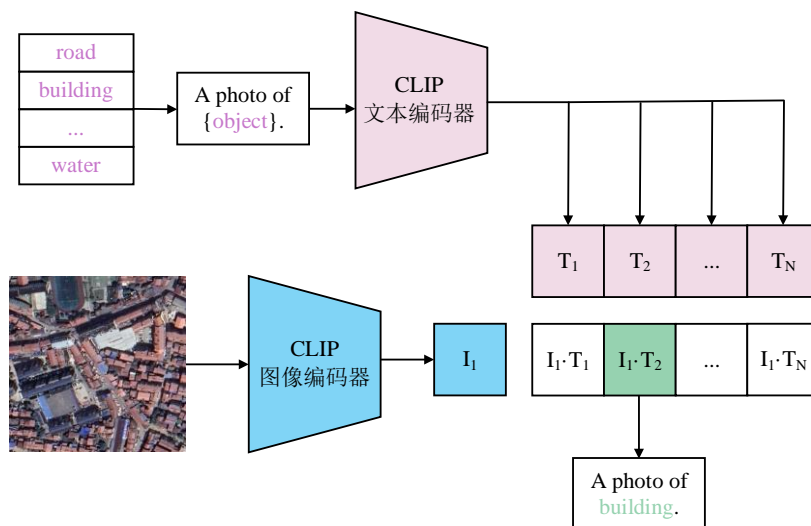


图 2.8 使用 CLIP 模型进行零样本推理

Fig.2.8 Zero-shot inference using the CLIP model

为了更充分地挖掘 CLIP 模型在图像-文本特征融合方面的潜力，成本体积被创新性地引入开放词汇语义分割任务中。成本体积（Cost Volume）是计算机视觉领域中用于立体匹配、深度估计及语义分割等任务的核心数据结构，用于表征不同假设（如视差或深度）下的匹配成本。在传统立体匹配中，成本体积通常表现为一个三维张量，其维度包括图像的高度、宽度以及视差假设数量，用于记录每个像素在不同视差下的匹配代价。而在多模态场景下，成本体积的概念被进一步扩展，以适应更复杂的任务需求。具体而言，在开放词汇语义分割中，成本体积通过计算 CLIP 模型生成的图像嵌入与文本嵌入之间的余弦相似度分数来构建，反映了图像像素与文本描述之间的语义匹配成本，因此，也可以将其称为语义匹配特征图。这种多模态语义匹配特征图突破了传统空间维度的限制，融入了丰富的语义信息，通过聚合图像与文本嵌入的相似性分数，形成高效的特征表示结构，从而实现对可见和不可见类别像素的精确分割。在 CAT-Seg 研究中，进一步探索了如何高效聚合这种多模态语义匹配特征图，例如采用空间聚合和类别聚合策略，以提升分割精度并优化计算效率^[61]。这种方法显著增强了模型对不可见类别的泛化能力，同时维持了对可见类别的出色分割性能。

2.4 遥感地学知识图谱（Remote sensing geoscience knowledge graph）

遥感地学知识图谱是为应对遥感图像分析挑战而设计的一种结构化知识表示形式，

通过节点和边系统化描述概念及其关系，展现出显著的应用潜力^[47]。它能捕捉地物类别的复杂语义关联和层次结构（如揭示“河流”与“湖泊”的相似性及“建筑物”与自然地物的差异），增强文本嵌入的语义表达，并通过编码节点属性和空间关系优化视觉特征表示，弥补传统方法在多维属性和空间分布捕捉上的不足。尤其在零样本任务中，知识图谱通过提供不可见类别的语义上下文和关系推理能力，显著提升模型的泛化性能，例如利用图谱中的关系推断未标记地物的语义属性或空间分布规律，从而实现对未知类别的有效识别。其数据来源包括：RSKG^[9]（遥感专用知识图谱，含 117 个实体、26 种关系、191 个三元组）和 ConceptNet^[49]（通用常识知识图谱，含 3855 个实体、15 种关系、3343 个三元组），两者结合为遥感任务提供专业性和通用性支持，特别在零样本场景下增强了跨领域知识迁移能力。

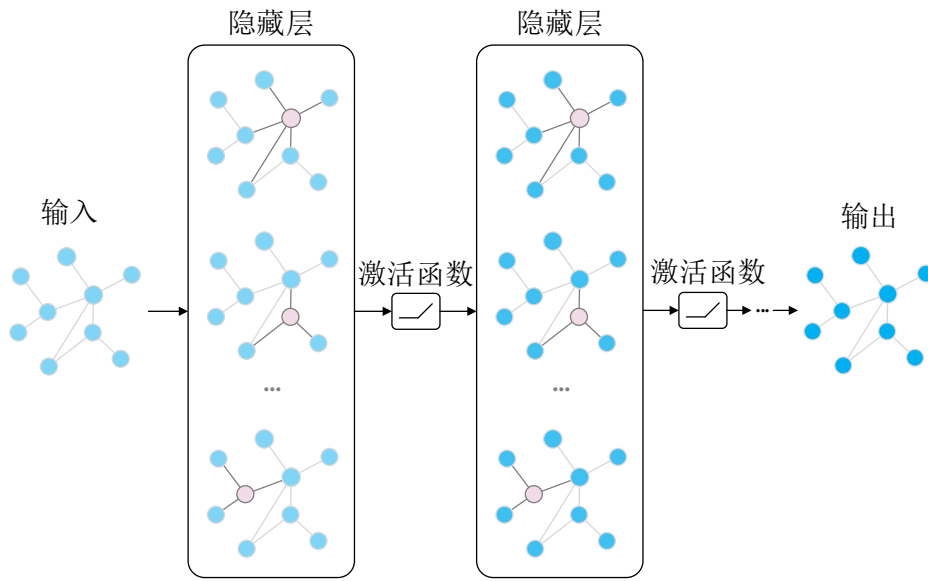


图 2.9 图卷积网络

Fig.2.9 Graph Convolutional Network

在遥感地学知识图谱的应用中，图卷积网络（Graph Convolutional Network, GCN）是一种关键技术，用于高效提取和建模图结构数据中的特征^[50]。GCN 通过对图的邻接矩阵和节点特征进行卷积操作，实现节点特征的更新和聚合，其核心公式为：

$$H^{(l+1)} = \sigma \left(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^{(l)} W^{(l)} \right) \quad (2.2)$$

其中 $H^{(l)}$ 表示第 l 层节点特征矩阵， $\tilde{A} = A + I$ 为带自环的邻接矩阵， \tilde{D} 为度矩阵， $W^{(l)}$ 为可学习的权重矩阵， σ 为非线性激活函数。通过这种方式，图卷积网络（GCN）能够聚合每个节点的邻域信息，有效捕捉图结构中的拓扑特性与语义关系。在遥感任务中，图卷积网络能够精确建模地物间的空间关联和语义依赖。例如，通过分析知识图谱中“河流”与“湖泊”之间的连接强度，可以显著提升对地物类别的判别精度。尤其在零样本学习场景下，GCN 能够利用图结构进行推理，生成不可见类别的特征表示，从而提高模型的泛化能力。这一方法为遥感领域中的复杂关系建模提供了有力的理论支持与

实践依据。

2.5 信息瓶颈理论 (Information bottleneck theory)

信息瓶颈理论 (Information Bottleneck Theory) 是一种用于理解信息处理和机器学习中数据压缩与特征提取的理论框架, 由 Naftali Tishby 等人于 1999 年提出^{[51][52]}。它旨在通过在保留任务相关信息的同时, 尽可能压缩输入数据, 从而实现一种最优的表示形式。这一理论的核心思想是找到一个中间表示, 既能最大化与输出 (目标变量) 的相关性, 又能最小化与输入数据的冗余信息。这种方法在深度学习、信号处理和统计建模等领域有着广泛的应用。

信息瓶颈理论的基本原理可以用信息论中的互信息 (Mutual Information) 来描述。假设有输入变量 X 和目标变量 Y , 目标是找到一个压缩表示 T , 使得 T 能够很好地预测 Y , 同时尽量减少对 X 的依赖。形式上, 信息瓶颈的目标是通过优化以下目标函数来实现:

$$\mathcal{L} = I(X; T) - \beta I(T; Y) \quad (2.3)$$

其中, $I(X; T)$ 表示输入 X 和压缩表示 T 之间的互信息, 衡量 T 保留了多少关于 X 的信息; $I(T; Y)$ 表示 T 和目标 Y 之间的互信息, 衡量 T 对 Y 的预测能力; β 是一个超参数, 用于权衡压缩和预测之间的取舍。当 β 较大时, 模型更倾向于保留与 Y 相关的信息, 而当 β 较小时, 则更倾向于压缩 X 。

在实际应用中, 信息瓶颈理论提供了一种解释深度神经网络行为的方式^[53]。例如, 它可以用来分析网络如何通过层级结构逐步丢弃与任务无关的信息, 同时保留关键特征。优化这一目标通常需要用到变分推理或近似方法, 因为直接计算互信息在高维数据中往往是不可行的。通过这种方式, 信息瓶颈理论不仅为信息处理提供了一个理论基础, 还为设计更高效的机器学习模型提供了启发。

2.6 零样本语义分割 (Zero-shot semantic segmentation)

零样本语义分割 (Zero-Shot Segmentation) 是指在无需针对特定类别进行训练的情况下, 准确分割图像中不可见类别的目标区域^[54]。相较于传统图像分割方法, 零样本语义分割突破了对大量标注数据的依赖。传统方法通常通过优化可见类别的标注数据实现分割, 但在实际应用中, 模型常需处理未见过的类别, 导致泛化能力受限。零样本语义分割通过引入外部知识, 如语义信息、知识图谱或预训练的多模态模型, 借助类别间的语义关系或跨模态信息 (如图像与文本的关联), 实现从可见类别到不可见类别的知识迁移, 从而显著提升模型的泛化能力。

例如, 基于 CLIP (Contrastive Language-Image Pretraining) 的零样本语义分割方法, 利用预训练的图像和文本编码器, 将图像像素与类别文本描述映射到统一的嵌入空间, 通过计算余弦相似度实现精准的像素级分类。这种方法无需针对新类别重新训练即可完成分割任务, 极大地降低了数据准备成本。此外, 知识图谱在零样本语义分割中也发挥

了重要作用。通过编码类别间的语义关系（如“树木”与“道路”的空间邻近性或功能关联性），知识图谱为模型提供了丰富的上下文信息，有助于推断不可见类别的语义特征。部分方法还结合生成模型（如扩散模型）或自监督学习，进一步增强不可见类别的特征表达能力。

在遥感图像分割领域，零样本语义分割具有尤为突出的应用价值。遥感图像中地物类别繁多（如建筑物、植被、水体等），且分布复杂，传统标注过程耗时且成本高昂。零样本语义分割通过利用外部知识和预训练模型，能够有效减少对标注数据的需求，同时提升模型对新场景和新类别的适应性。例如，在城市规划、灾害监测或农业管理中，模型可快速识别未标注的稀有地物类别，从而支持实时决策。

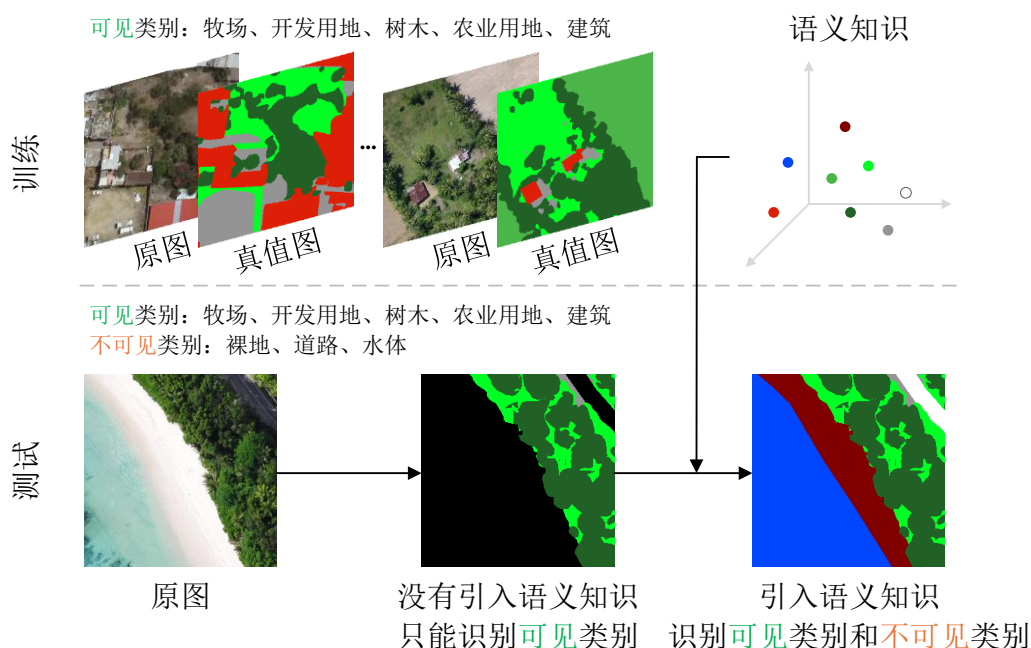


图 2.10 零样本语义分割

Fig.2.10 Zero-Shot Semantic Segmentation

在零样本语义分割框架下，可区分为广义零样本语义分割（Generalized Zero-Shot Semantic Segmentation, GZLSS）和狭义零样本语义分割（Zero-Shot Semantic Segmentation, ZSS）。广义零样本语义分割要求测试阶段同时分割可见类别 C_s 和不可见类别 C_u ，其中 $C_s \cap C_u = \emptyset$ ，而狭义仅要求分割不可见类别 C_u 。广义设置更具挑战性，需平衡可见与不可见类别的分割精度，避免对可见类别的偏向。本文提出的方法遵循广义零样本语义分割（GZLSS），仅使用可见类像素标注数据训练，训练阶段根据可见类的语义描述生成像素分类结果，测试阶段对可见和不可见类均产生分割结果，且 $C_s \cap C_u = \emptyset$ ，训练时不可见类别标签不可用。

零样本语义分割方法大致可以分为两阶段方法和一阶段方法。两阶段方法将掩码生成与掩码分类解耦，首先通过掩码提议网络生成类别无关的掩码，随后利用预训练的图像-文本模型（如 CLIP）对这些掩码区域逐一分类。这种方法有效保留了 CLIP 的零样本

能力，但因需对图像中大量掩码区域单独分类，导致推理成本较高。相比之下，一阶段方法通过直接微调预训练的文本-图像模型，或将其知识蒸馏至辅助网络上应用于分割数据集，显著降低了推理成本。然而，这一过程可能牺牲了预训练模型的部分图像-文本对齐能力。两者在性能与效率间各有取舍，适于不同应用场景。

2.7 本章小结 (Summary)

本章系统探讨了深度学习领域的核心技术。首先，详细介绍了卷积神经网络的架构，阐述了其关键组成部分——卷积层、池化层和激活层的作用与原理，并简要概述了常用的图像特征提取主干网络。其次，深入分析了遥感地学知识图谱的结构特性，以及图卷积网络作为图结构建模工具的应用与优势。接着，介绍了信息瓶颈理论及其应用。最后，对零样本技术进行了全面介绍，突出了其在解决不可见类别问题中的重要价值与应用潜力。

3 多尺度增强和旋转不变性零样本语义分割方法

3 Multi-scale Enhancement and Rotation Invariance-based Zero-shot Segmentation Method

3.1 引言 (Foreword)

遥感图像零样本语义分割作为计算机视觉领域的重要研究方向，在城市规划、环境监测和农业管理等多个领域展现出广阔的应用潜力。遥感图像通过卫星或无人机等设备采集，具有覆盖范围广、数据量大以及地表信息丰富的特点。在零样本学习场景下，即在缺乏标注数据的情况下实现语义分割，近年来已有若干研究尝试解决相关问题。例如，Ding 等人提出 ZegFormer 模型，通过将零样本语义分割解耦为类无关像素分组与片段级分类的双任务框架，构建像素-片段层级联处理结构，并引入 CLIP 的视觉语言模型实现片段级语义推理^[12]；Xu 等人构建了两阶段开放词汇分割基线，采用掩码提议生成与图像裁剪分类的协同架构，通过分离区域定位与语义推理过程解决 CLIP 图像-像素粒度差异问题^[58]；Xu 等人提出侧适配网络 (SAN) 框架，通过将冻结 CLIP 模型与双分支侧网络（掩码提议生成分支、注意力偏置预测分支）进行级联，构建掩码区域识别与语义分类解耦的端到端架构，利用注意力偏置动态调整 CLIP 特征以实现掩码语义的 CLIP 自适应表征^[59]。尽管这些工作在零样本语义分割领域取得了一定进展，但它们在应用于遥感图像零样本语义分割时存在局限性。首先，这些方法主要针对自然图像设计，未充分考虑遥感图像中目标尺度的显著差异性，导致其特征提取难以有效表征跨尺度的语义信息。其次，这些方法在处理遥感图像多视角成像特性时表现不足，目标可能以任意角度呈现，而这些方法缺乏对旋转不变性的优化，限制了模型的鲁棒性。此外，由于遥感图像场景复杂且缺乏标注数据支持，CLIP 的跨模态对齐能力难以直接挖掘类别间的语义关系，影响模型在新场景或新类别上的泛化能力。

为应对遥感图像零样本语义分割中的关键难题，本文提出了三种核心模块，分别针对多尺度特征表达、旋转不变性和类别关系建模进行优化。首先，全局-局部信息增强模块 (Global-Local Information Enhancement Module, GLIEM) 通过融合 CLIP 的全局语义特征与 ResNet 的局部细节特征，结合特征金字塔网络 (FPN) 实现跨尺度特征的有效整合，从而构建包含丰富多尺度信息的特征表示，克服目标尺度差异带来的挑战。其次，旋转不变性特征优化模块 (Rotation Invariance Feature Optimization Module, RIFOM) 针对多视角成像特性，通过多角度旋转生成等变特征，并结合池化操作和空间注意力机制提取旋转不变特征，从而提升模型对目标角度变化的鲁棒性。最后，类别关系建模模块 (Category Relationship Modeling Module, CRMM) 利用自注意力机制分析遥感图像中类别

间的语义关联，使各类别特征能够相互关联，从而提升模型对复杂场景的分割能力。

3.2 多尺度增强和旋转不变性网络 MSRI Net (Multi-scale enhancement and rotational invariance network)

3.2.1 MSRI Net 整体架构

MSRI Net 的整体架构如图 3.1 所示。考虑到 CLIP 模型的 Vision Transformer (ViT) 视觉编码器主要提取全局语义信息，而遥感图像中局部细节特征对于实现精准的语义分割至关重要，因此引入预训练的 ResNet 模型以捕获图像的局部细节特征。具体而言，遥感图像首先分别输入至 CLIP 的 ViT 视觉编码器和预训练 ResNet 模型，生成各自的多尺度特征图。随后，这两组特征图被输入至全局-局部信息增强模块 (GLIEM)，以生成融合全局-局部信息的视觉特征图。GLIEM 模块通过整合 CLIP 视觉编码器提取的浅层全局特征与 ResNet 提取的局部细节特征，并利用特征金字塔网络 (FPN) 实现全局-局部特征的融合，构建出融合了多尺度信息的视觉特征图。接着，融合后的视觉特征图经过特征重排为视觉嵌入，与 CLIP 模型的文本嵌入通过余弦相似度计算，生成语义匹配特征体。该特征体表征了图像局部区域与各类别文本描述之间的语义匹配程度的空间分布，可以被视为粗略的语义掩码。此后，语义匹配特征体依次输入至旋转不变性特征优化模块 (RIFOM) 和类别关系建模模块 (CRMM) 进行进一步优化。在 RIFOM 模块中，通过多角度旋转操作生成等变特征，并结合空间注意力机制提取旋转不变特征，以增强模型对多视角目标的适应能力；在 CRMM 模块中，采用自注意力机制挖掘类别间的复杂语义关系，进一步优化各类别的语义表征。最后，经过 RIFOM 和 CRMM 优化后的特征输入至分割头，生成最终的零样本语义分割结果。

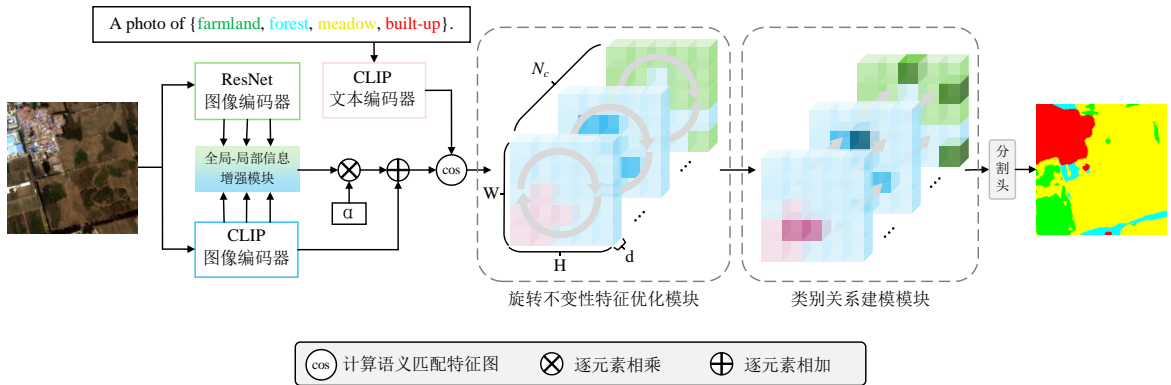


图 3.1 MSRI Net 整体架构

Fig.3.1 The overall architecture of MSRI Net

3.2.2 全局-局部信息增强模块

针对遥感图像的多尺度特性，本文提出了一种全局-局部信息增强模块 (GLIEM)，旨在优化 CLIP 模型输出的视觉特征，使其更好地适应遥感图像零样本语义分割任务。鉴

于 CLIP 模型所采用的 Vision Transformer (ViT) 视觉编码器侧重于提取图像的全局语义表征, 而遥感影像中地物边界、纹理模式等局部细节对于实现精准的语义分割至关重要。通过引入预训练的 ResNet-50 模型作为辅助特征提取器, 利用其卷积神经网络的特性, 在多个尺度上捕获局部细节特征, 从而形成全局语义引导与局部细节增强的特征互补机制。具体而言, 该模块通过结合 CLIP 模型的全局语义表示能力与 ResNet 预训练模型的局部细节特征提取能力, 构建包含全局-局部信息的特征表示, 并进一步通过特征金字塔网络 (FPN) 实现分层递进式融合, 从而提升模型对跨尺度目标的表征能力。

对于输入的遥感图像 I , 通过 CLIP 的视觉编码器 $\Phi^V(\cdot)$ 对图像进行特征提取, 得到密集的视觉嵌入 $E^V = \Phi^V(I) \in \mathbb{R}^{H \times W \times d}$ 。其中, H 为特征图的高度, W 为特征图的宽度, d 表示嵌入维度。为了在保留 CLIP 视觉编码器全局语义信息的同时, 更好地融合局部细节信息, 我们提取了 CLIP 视觉编码器的浅层特征 (第 1、3、5 层), 分别记为 $F_{\text{clip}}^1, F_{\text{clip}}^3, F_{\text{clip}}^5$ 。这些浅层特征能够提供局部细节, 有助于后续的特征融合。为适配后续特征融合的空间分辨率需求, 利用上采样操作对上述特征进行空间分辨率调整, 并用 1×1 卷积进行通道数调整, 生成 CLIP 多尺度特征集合 $\mathcal{F}_{\text{clip}}$:

$$\mathcal{F}_{\text{clip}} = \{\text{Conv}_{1 \times 1}(\text{Up}_{192}(F_{\text{clip}}^1)), \text{Conv}_{1 \times 1}(\text{Up}_{96}(F_{\text{clip}}^3)), \text{Conv}_{1 \times 1}(\text{Up}_{48}(F_{\text{clip}}^5))\} \quad (3.1)$$

其中, Up_i 表示针对不同尺度特征的上采样操作, i 表示目标分辨率。

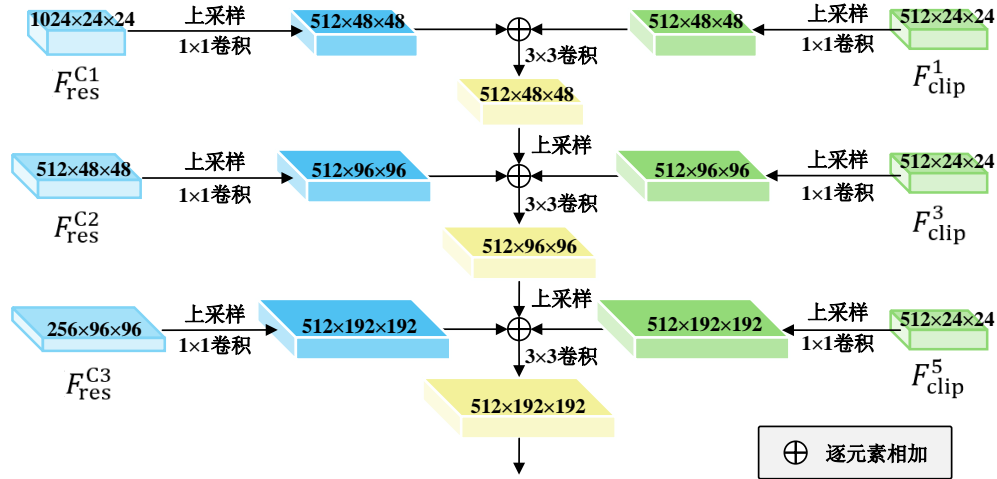


图 3.2 全局-局部信息增强模块

Fig.3.2 Global-Local Information Enhancement Module

为充分利用 ResNet 预训练模型的局部细节提取能力, 本文通过 ResNet 模型提取输入图像 I 的多尺度局部细节特征。具体地, 从 ResNet 的中间层 (C1、C2、C3 层) 提取特征 $F_{\text{res}}^{C1}, F_{\text{res}}^{C2}, F_{\text{res}}^{C3}$, 这些特征能够捕捉图像中的局部细节信息, 为后续的特征融合提供细节特征支持。为了适配后续特征的融合, 利用上采样操作对 ResNet 特征进行分辨率调整, 使其与 $\mathcal{F}_{\text{clip}}$ 的特征维度保持一致, 并用 1×1 卷积进行通道数调整, 最终生成对齐后的

ResNet 提取的多尺度特征集合 \mathcal{F}_{res} :

$$\mathcal{F}_{\text{res}} = \{\text{Conv}_{1 \times 1}(\text{Up}_{192}(F_{\text{res}}^{C1})), \\ \text{Conv}_{1 \times 1}(\text{Up}_{96}(F_{\text{res}}^{C2})), \\ \text{Conv}_{1 \times 1}(\text{Up}_{48}(F_{\text{res}}^{C3}))\} \quad (3.2)$$

为实现全局-局部特征的有效融合, 本文采用特征金字塔网络 (FPN) 对 \mathcal{F}_{res} 和 $\mathcal{F}_{\text{clip}}$ 进行融合。具体地, 将 \mathcal{F}_{res} 和 $\mathcal{F}_{\text{clip}}$ 进行逐元素相加, 并通过 FPN 网络进行分层递进式融合, 得到全局-局部融合特征 \mathcal{F}_{GL} :

$$\mathcal{F}_{\text{GL}} = \text{FPN}(\mathcal{F}_{\text{res}} \oplus \mathcal{F}_{\text{clip}}) \quad (3.3)$$

其中, \mathcal{F}_{GL} 包含多个尺度的融合特征, 分别对应不同的空间分辨率。

为提高 CLIP 的全局-局部表达能力, 利用 FPN 输出的最高分辨率特征对原始 CLIP 视觉特征进行增强。具体地, 将 FPN 输出的最高分辨率特征图通过上采样和投影操作调整至与 CLIP 视觉嵌入 E^V 相同的维度, 得到全局-局部增强嵌入 E_{GL} , 随后, 通过自学习的加权重 α 进行加权融合得到全局-局部增强视觉嵌入 E_{GL}^V :

$$E_{\text{GL}}^V = E^V + \alpha \cdot E_{\text{GL}} \quad (3.4)$$

为了同时兼顾图像的视觉信息和文本语义信息, 以更好地实现遥感场景中的零样本语义分割, 通过计算语义匹配特征图来辅助分割。给定一组类别 c , 首先通过 CLIP 文本编码器 $\Phi^L(\cdot)$ 提取出类别文本嵌入 $E^L = \Phi^L(c) \in \mathbb{R}^{N_c \times d}$, 其中 N_c 为类别数。接下来, 利用文本嵌入 E^L 和全局-局部增强视觉嵌入 E_{GL}^V 进行余弦相似度计算, 从而计算语义匹配特征图 $F_{\text{cos}} \in \mathbb{R}^{H \times W \times N_c}$:

$$F_{\text{cos}} = \frac{E_{\text{GL}}^V \cdot E^L}{\|E_{\text{GL}}^V\| \|E^L\|} \quad (3.5)$$

语义匹配特征图能够有效捕捉图像局部视觉特征与类别语义嵌入之间的匹配度, 可以被视为每个类别的粗略语义掩码, 从而为后续的零样本语义分割提供更为精确的依据。为了进一步的增强语义匹配特征图在高维空间的处理以适应精细的分割任务, 在类别维度上使用卷积层进行处理得到语义匹配特征体 $F_{\text{COS}} \in \mathbb{R}^{H \times W \times N_c \times d}$, 其中 d 是通道数。

3.2.3 旋转不变性特征优化模块

旋转不变性是指特征表达在目标发生任意角度旋转后仍能保持一致的能力, 这一特性在处理遥感图像时尤为关键。遥感图像因多视角拍摄特性, 目标常以任意角度呈现, 这对特征提取提出了保留空间上下文信息并实现旋转不变的需求。为此, 本文提出了一种旋转不变性特征优化模块, 旨在通过空间维度的特征优化解决这一挑战。该模块的核心设计在于: 首先通过多角度旋转生成等变特征, 随后结合池化操作提取旋转不变特征, 最后引入空间注意力机制增强关键区域的表达, 从而在支持后续任务的同时兼顾空间上下文信息的完整性。

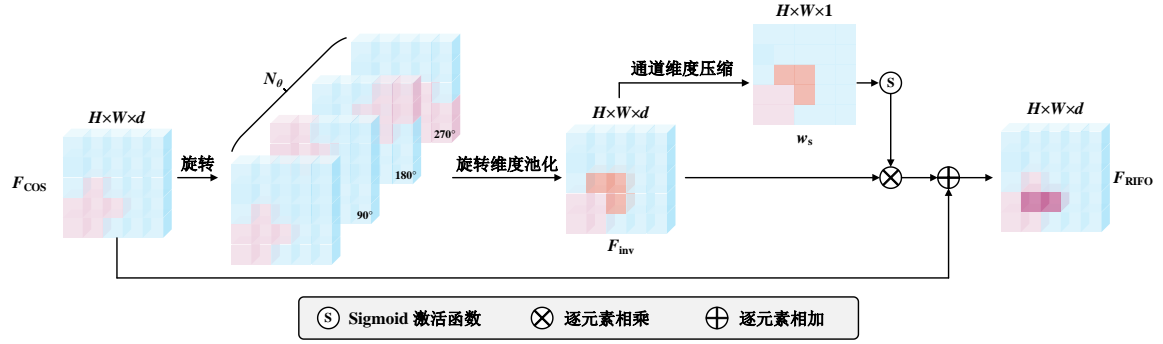


图 3.3 旋转不变性特征优化模块

Fig.3.3 Rotation Invariance Feature Optimization Module

RIFOM 的输入为语义匹配特征体 $F_{\text{COS}} \in \mathbb{R}^{H \times W \times N_c \times d}$ ，其中 d 表示通道数， N_c 表示类别数， H 和 W 分别表示特征图的高度和宽度。首先对语义匹配特征体 F_{COS} 在空间维度上进行多角度旋转，定义旋转角度集合 $\theta = \{\theta_i \mid \theta_i = i \cdot 360^\circ / N_\theta, i = 0, 1, \dots, N_\theta - 1\}$ ，其中 N_θ 是旋转角度数量，对于每个角度 θ_i ，通过旋转操作生成旋转等变特征图 $F_{\theta_i} = \text{Rotate}(F_{\text{COS}}, \theta_i)$ ，并获取旋转后的特征图集合 $\{F_{\theta_i}\}_{i=0}^{N_\theta-1} \in \mathbb{R}^{H \times W \times N_c \times d \times N_\theta}$ ；随后对该集合沿旋转维度 N_θ 进行最大池化操作提取旋转不变特征，得到 $F_{\text{inv}} \in \mathbb{R}^{H \times W \times N_c \times d}$ ，公式如下：

$$F_{\text{inv}} = \text{MaxPool}(\{F_{\theta_i}\}_{i=0}^{N_\theta-1}) \quad (3.6)$$

接着为增强关键区域表达能力，对 F_{inv} 引入空间注意力机制，沿通道维度进行平均池化后通过 1×1 卷积、批归一化和 Sigmoid 激活函数生成空间权重描述符 $w_s \in \mathbb{R}^{H \times W \times N_c}$ ，公式如下：

$$w_s = \sigma \left(\text{BN} \left(\text{Conv}_{1 \times 1} (\text{AvgPool}(F_{\text{inv}})) \right) \right) \quad (3.7)$$

并通过加权操作得到空间增强特征 $F_s = F_{\text{inv}} \cdot w_s$ ；随后，为保持通道数一致并提升训练稳定性，将 F_s 通过 1×1 卷积调整通道数并恢复形状后，结合残差连接融合语义匹配特征体 F_{COS} ，得到最终输出 $F_{\text{RIFO}} \in \mathbb{R}^{H \times W \times N_c \times d}$ ，公式如下：

$$F_{\text{RIFO}} = \text{Conv}_{1 \times 1}(F_s) + F_{\text{COS}} \quad (3.8)$$

3.2.4 类别关系建模模块

类别关系建模旨在挖掘并表征数据中不同类别之间的语义关联，以提升特征表达的针对性和鲁棒性。在遥感图像中，各类别（如建筑物、农田、河流等）往往呈现复杂的语义关系，例如建筑物多集中于城市区域，而农田常与河流邻近。为充分利用这些关系并增强关键类别的特征表达，本文提出了一种类别关系建模模块（CRMM）。该模块通过自注意力机制捕捉类别间的语义依赖关系，从而优化特征表示以适应遥感图像的复杂性。

CRMM 的输入为经过旋转不变性特征优化模块的特征体 $F_{\text{RIFO}} \in \mathbb{R}^{H \times W \times N_c \times d}$ 。实现过程首先将 F_{RIFO} 通过 1×1 卷积降维来降低计算成本，得到轻量级特征体 $F_{\text{RIFO}_s} \in$

$\mathbb{R}^{(H \times W) \times N_c \times d_r}$, 公式如下:

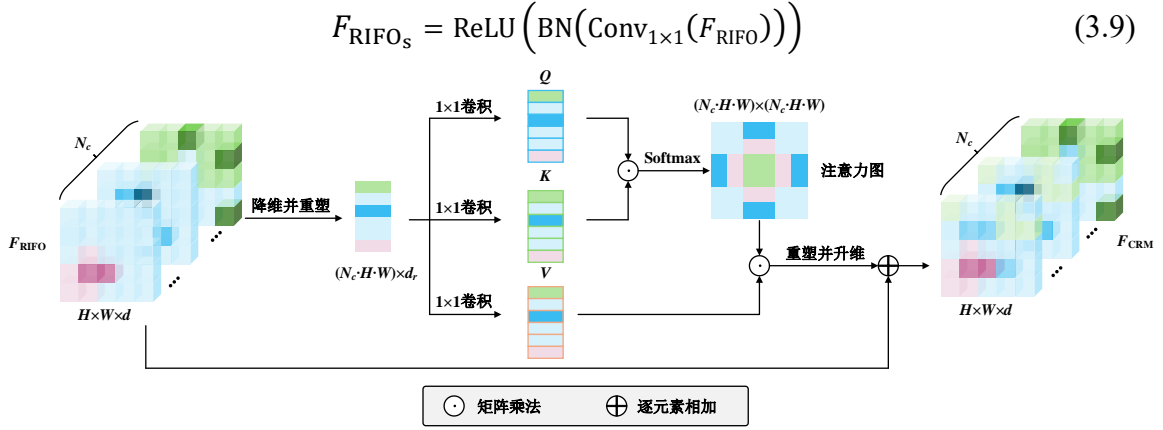


图 3.4 类别关系建模模块

Fig.3.4 Category Relationship Modeling Module

随后采用自注意力机制挖掘类别间的语义关系, 为了实现类别间的交互, 将类别维度 N_c 和空间维度 $H \times W$ 合并为一个维度, 使得一个类别的特征能够关注所有类别的特征。具体地, 将 F_{RIFO_s} 通过三个 1×1 卷积生成 Query、Key 和 Value 向量 $Q = \text{Conv}_{1 \times 1}^Q(F_{\text{RIFO}_s})$, $K = \text{Conv}_{1 \times 1}^K(F_{\text{RIFO}_s})$, $V = \text{Conv}_{1 \times 1}^V(F_{\text{RIFO}_s})$, 其中 $Q, K, V \in \mathbb{R}^{(N_c \cdot H \cdot W) \times d_r}$ 。接着, 计算加权类别特征 $F_w \in \mathbb{R}^{N_c \times d_r \times (H \cdot W)}$, 计算公式如下:

$$F_w = \text{Softmax} \left(\frac{Q^T K}{\sqrt{N_c \cdot H \cdot W}} \right) V^T \quad (3.10)$$

其中 $QK^T \in \mathbb{R}^{(N_c \cdot H \cdot W) \times (N_c \cdot H \cdot W)}$, 表示个类别的每个空间位置可以关注所有类别的所有空间位置, 从而实现类别间的语义交互。缩放因子 $\sqrt{N_c \cdot H \cdot W}$ 用于防止数值过大。

随后, 为保持通道数一致并提升训练稳定性, 通过 1×1 卷积调整通道数, 从 d_r 恢复到 d 。然后, 结合残差连接融合原始特征 F_{RIFO} , 得到最终输出 $F_{\text{CRM}} \in \mathbb{R}^{H \times W \times N_c \times d}$, 公式如下:

$$F_{\text{CRM}} = \text{Conv}_{1 \times 1}(F_w) + F_{\text{RIFO}} \quad (3.11)$$

最后使用一个简单的二维卷积将 d 维映射到 1 维并上采样到原图分辨率, 生成最终的分割预测结果。MSRINet 的流程在算法 1 中进行了详细阐述。

3.3 实验结果与分析 (Experimental results and analysis)

3.3.1 实验环境

本章使用 PyTorch 深度学习框架实现深度学习模型, 基于 NVIDIA A100 显卡进行高效模型训练, 采用交叉熵损失作为优化目标, 配置包括每批次图像数为 4, 学习率调度器为 WarmupCosineLR, 基础学习率设为 0.0002, 最大迭代次数为 30000, 梯度裁剪倍数为 0.01, 并在训练过程中每 5000 次迭代进行一次评估, 以确保模型性能的稳定提升。

 算法 1: 基于多尺度增强和旋转不变性的零样本语义分割

输入:

- I_s : 仅含可见类别的遥感图像 (训练时使用)
- I_u : 含不可见类别的遥感图像 (测试时使用)
- C_s : 可见类别
- C_u : 不可见类别
- GT_s : 可见类别的像素级标注 (训练时使用, 仅对 C_s 对应的类别)

输出:

- S : 输入图像的语义分割掩码 (包含不可见类别)

算法:

1. 训练阶段 (仅针对可见类别):
 - 1) 输入训练图像 I_s 和对应的可见类别标注 GT_s 。
 - 2) 使用 CLIP 视觉编码器从 I_s 提取多尺度特征集合 \mathcal{F}_{clip} 。
 - 3) 使用 ResNet 从 I_s 提取多尺度特征集合 \mathcal{F}_{res} 。
 - 4) 全局-局部信息增强模块 (GLIEM):
 - a) 通过特征金字塔网络融合 \mathcal{F}_{clip} 和 \mathcal{F}_{res} , 生成全局-局部融合特征 \mathcal{F}_{GL} 。
 - b) 将 \mathcal{F}_{GL} 的最高分辨率特征图通过上采样和投影得到全局-局部增强嵌入 E_{GL} 。
 - c) 将 E_{GL} 和 CLIP 视觉编码器的视觉嵌入 E^V 进行加权求和得到全局-局部增强视觉嵌入 E_{GL}^V 。
 - d) 将 E_{MS} 和使用 CLIP 文本编码器从 C_s 生成的可见类别嵌入 E^L 进行余弦相似度计算并卷积, 从而获得语义匹配特征体 F_{COS} 。
 - 5) 旋转不变性特征优化模块 (RIFOM):
 - a) 对 F_{COS} 进行多角度旋转和池化, 生成旋转不变特征 F_{inv} 。
 - b) 使用空间注意力机制获得 F_s 。
 - c) 使用残差连接获得经过 RIFOM 优化的 F_{RIFO} 。
 - 6) 类别关系建模模块 (CRMM):
 - a) 对 F_{RIFO} 进行降维获得 F_{RIFO_s} 。
 - b) 使用自注意力挖掘 F_{RIFO_s} 之间的语义关系获得加权类别特征 F_w 。
 - c) 使用残差连接获得经过 CRMM 优化的 F_{CRM} 。
 - 7) 使用 GT_s 监督训练, 优化模型参数。
 2. 测试阶段 (针对所有类别):
 - 1) 输入测试图像 I_u 、可见类别 C_s 和不可见类别 C_u 。
 - 2) 经过 MSRI Net 推理生成分割掩码 S 。
 3. 返回 S
-

3.3.2 实验数据集

(1) GID5 数据集

GID5 数据集 (Gaofen Image Dataset) 是一个基于中国高分系列卫星影像构建的高分辨率遥感数据集, 专为土地覆盖分类任务设计, 广泛应用于环境监测与城市规划研究^[55]。它包含数千张覆盖中国多种地理区域 (如城市、农村、山地等) 的图像, 分辨率高达 0.8 米至 2 米, 提供 5 个主要类别的高质量像素级标注, 包括建筑区、农田、森林、草地、水体。该数据集的影像经过精细校正, 具备多光谱特性, 支持多任务学习, 例如语义分割、变化检测和目标识别。GID5 的多样性和高分辨率特性使其成为测试深度学习模型鲁棒性和精度的理想选择, 尤其在处理复杂地貌和细粒度分类时表现突出。

表 3.1 实验遥感数据集介绍

Table 3.1 Introduction to Experimental Remote Sensing Datasets

数据集名称	GID5	OpenEarthMap	LoveDA
数据来源	中国高分二号卫星 (GF-2)	多源 (xBD、Inria、SpaceNet 等 + 公开航拍数据)	中国高分卫星 + Google Earth (南京、常州、武汉)
分辨率	0.8m (PAN) / 3.24m (MS)	0.25–0.5m GSD	0.3m
覆盖范围	中国 (50,000 km ² , 60+城市)	全球 (44 国, 97 区域, 6 大洲)	中国 3 个城市 (536 km ²)
图像数量	150 张 (6800×7200 像素)	5000 张 (1024×1024 像素)	5987 张 (1024×1024 像素)
标注类别	建筑区、农田、森林、草地、水体	裸地、牧场、开发用地、道路、树木、水体、农业用地、建筑	背景、建筑、道路、水、裸地、森林、农田
标注特点	像素级标注, 侧重农业和城市用地	精细化标注, 覆盖全球多样性地貌	区分城乡域, 强调跨域适应性
独特优势	中国区域高覆盖、多时相数据	全球代表性、高密度标注	首个针对城乡域适应的遥感数据集

(2) OpenEarthMap 数据集

OpenEarthMap 是一个全球范围的高分辨率土地覆盖映射基准数据集, 旨在为遥感领域的开源研究提供标准化资源, 推动人工智能在地球观测中的应用^[56]。它收集了来自 44 个国家、97 个代表性地区的 5000 张航空和卫星图像, 覆盖亚洲、欧洲、北美、南美、非洲和大洋洲六大洲, 地面采样距离 (GSD) 介于 0.25 米至 0.5 米之间。数据集标注了 8 个土地覆盖类别, 包括裸地、牧场、开发用地、道路、树木、水体、农业用地、建筑, 每张图像尺寸为 1024×1024 像素, 并提供 RGB 三通道数据。OpenEarthMap 的全球分布特性使其训练的模型具备跨区域适应性, 适用于气候变化分析、灾害评估和土地使用规划等场景, 同时其开源性质也降低了研究门槛, 促进了学术界的合作与创新。

(3) LoveDA 数据集

LoveDA (Land-cover Domain Adaptive) 数据集是一个专为遥感土地覆盖语义分割和域自适应任务设计的高质量数据集, 包含 5987 张分辨率为 0.3 米的遥感图像, 数据来源于中国南京、常州和武汉三个城市, 涵盖城市和农村两种典型地理环境。该数据集标注了 7 个类别——建筑、道路、水体、裸地、森林、农业用地和背景, 总计超过 25 万多边形标

注，特别突出了多尺度目标（从小型房屋到大型农田）、复杂背景（如阴影、植被遮挡）和类别分布不均（如城市建筑密集 vs. 农村农田稀疏）三大挑战^[57]。LoveDA 不仅适用于传统的监督式语义分割任务，还为无监督域自适应研究提供了丰富的实验场景，能够有效评估模型在不同地域间的泛化能力。其设计初衷是推动遥感技术在精细化土地覆盖映射中的应用，助力智慧城市建设、农业管理及生态保护等领域的发展。

3.3.3 零样本数据集制作

为开展零样本语义分割研究，本文基于 GID5、OpenEarthMap 和 LoveDA 三个遥感数据集，精心构建了适用于零样本学习的专用数据集。零样本学习要求训练阶段仅依赖可见类别的图像和标注数据，而不可见类别则留待测试阶段以检验模型的泛化性能。为满足这一要求，对原始数据集进行了系统化的筛选与处理，确保数据能够支持零样本任务的特殊需求。表 3.1 总结了各数据集在裁剪和筛选后的信息。

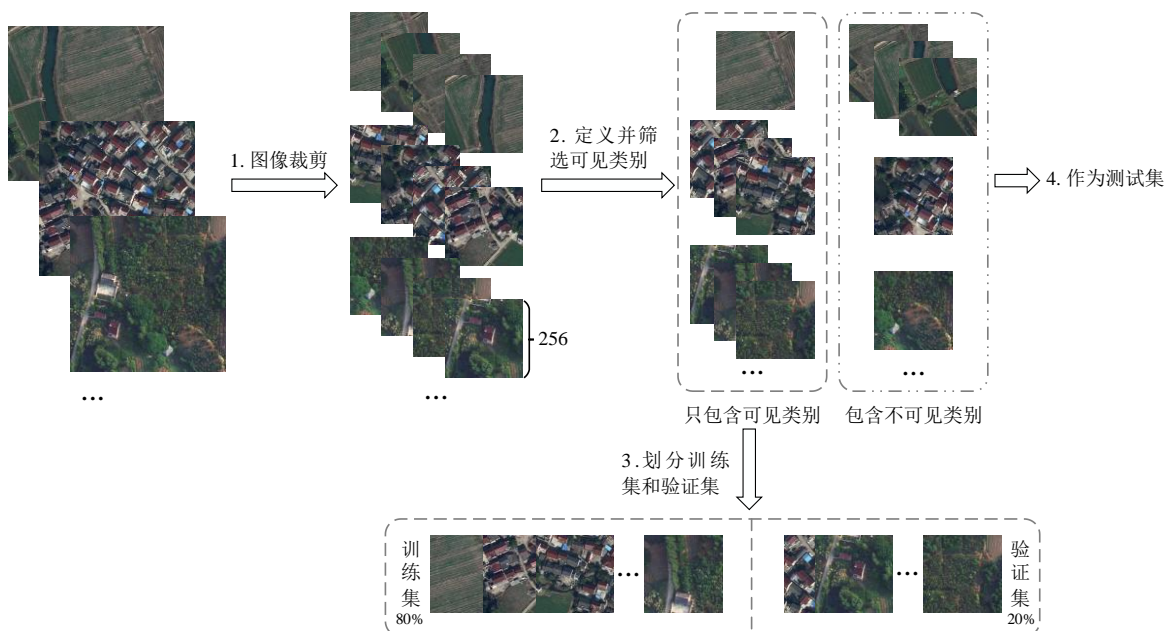


图 3.5 零样本语义分割数据集制作流程

Fig.3.5 The process of making a zero-shot semantic segmentation dataset

（1）图像裁剪

为尽可能增加训练样本数量，将三个数据集中的原始遥感图像裁剪为互不重叠的 256×256 像素小块。这一尺寸既能保留图像中的关键语义信息（如建筑轮廓、道路结构等），又能适配深度学习模型的批量处理需求，从而提升计算效率并为后续筛选提供充足的样本基础。

（2）定义可见类别和不可见类别

根据零样本学习的核心约束——训练阶段仅使用可见类别的图像和标签，并在测试阶段同时评估对于可见类别和不可见类别的识别能力，为每个数据集分别设计了可见类别和不可见类别的划分方案。

对于 GID5 数据集，从建筑区、农田、森林、草地、水体中选取建筑区、农田、水体作为可见类别，而将森林和草地设为不可见类别，以增加植被相关类别的测试难度。

对于 OpenEarthMap 数据集，从裸地、牧场、开发用地、道路、树木、水体、农业用地、建筑中选择牧场、开发用地、树木、农业用地、建筑作为可见类别，剔除裸地、道路和水体作为不可见类别，旨在突出模型对道路和水域的零样本预测能力，同时兼顾类别多样性与任务挑战性。

对于 LoveDA 数据集，从建筑、道路、水体、裸地、森林、农业用地中选取建筑、道路、森林、农业用地作为可见类别，将水体和裸地指定为不可见类别，以模拟自然环境和水域特征的部分缺失场景。

（3）筛选可见类别图像块

在裁剪后的图像块基础上，实施了严格的筛选流程，仅保留满足以下条件的样本：图像块中至少包含一种可见类别的像素，且完全不包含任何不可见类别的像素。这一筛选过程通过逐像素检查标注数据完成，确保训练样本的纯净性，严格符合零样本学习的约束条件，从而为模型训练提供高质量的数据支持。

表 3.2 零样本遥感数据集统计表格

数据集	总文件数	有效文件数	筛选比例	可见类别	不可见类别
GID5	117450	82244	70.02%	建筑区、农田、水体	森林、草地
OpenEarthMap	54138	12883	23.80%	牧场、开发用地、树木、农业用地、建筑	裸地、道路、水体
LoveDA	67056	38667	57.66%	建筑、道路、森林、农业用地	水体、裸地

（4）划分训练集和验证集

从筛选出的有效图像块中，按照 8:2 的比例随机划分训练集和验证集。其中，80%的样本构成训练集，用于优化模型参数；20%的样本作为验证集，用于调整超参数并监控训练过程中的性能表现。这一划分方式旨在充分利用有效数据，同时保证评估的客观性。

（5）确定测试集

筛选流程中被剔除的图像块（即包含不可见类别的样本）被保留作为测试集。这些数据专门用于评估模型在零样本场景下的泛化能力，测试其对不可见类别的预测性能，从而验证模型的鲁棒性和适应性。

3.3.4 评价指标

在本实验中，为了全面评估模型的性能，选用了四种常用的评估指标：交并比（IoU）、准确率（ACC）、均值交并比（mIoU）以及均值准确率（mACC）。这些指标能够从不同角度量化模型在目标检测与分割任务中的表现，具有广泛的代表性。具体而言，IoU 用于衡量预测区域与真实区域的重叠程度，其计算公式为：

$$IoU = \frac{|A \cap B|}{|A \cup B|} \quad (3.12)$$

其中 A 表示预测区域, B 表示真实区域, $|A \cap B|$ 和 $|A \cup B|$ 分别表示交集和并集的面积。ACC 表示模型分类的整体正确率, 计算公式为:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (3.13)$$

其中 TP 、 TN 、 FP 、 FN 分别代表真阳性、真阴性、假阳性和假阴性样本的数量。为了进一步评估模型在多类别任务中的表现, 引入了 $mIoU$ 和 $mACC$, 其中 $mIoU$ 定义为所有类别 IoU 的平均值, 计算公式为:

$$mIoU = \frac{1}{N_c} \sum_{i=1}^N IoU_i \quad (3.14)$$

N_c 为类别数; $mACC$ 则是各类别准确率的平均值, 计算公式为:

$$mACC = \frac{1}{N_c} \sum_{i=1}^N \frac{TP_i}{TP_i + FP_i + FN_i} \quad (3.15)$$

3.3.5 对比实验分析

为了全面验证 MSRINet 在遥感图像处理任务中的性能与有效性, 本文基于 GID5、OpenEarthMap 和 LoveDA 等多个具有代表性的公开数据集, 设计了系统性的对比实验, 与一系列先进的零样本分割方法进行了深入比较, 分析了其在不同地貌特征和复杂场景下的鲁棒性与泛化能力。其中包括两阶段的方法 ZegFormer^[12](CVPR2022)、zsseg^[58](ECCV2022)、SAN^[59](CVPR2023)和计算语义匹配特征图的 SED^[60](CVPR2024)、Cat-Seg^[61](CVPR2024)方法。

图 3.6 展示了不同方法在 GID5 数据集上的预测结果, 可视化结果表明, MSRINet 在遥感图像语义分割任务中表现优异, 尤其是在不可见类别(森林和草地)的分割上。相比其他方法, MSRINet 能够更精确地识别森林和草地的边界, 显著减少与其他类别的混淆, 这得益于其全局-局部信息增强模块和旋转不变性特征优化模块的强大适应性。ZegFormer 和 zsseg 在不可见类别上表现较弱, 例如森林常被误判为农田, 甚至对可见类别(如水体)的区分也不够精准。SAN 虽高效轻量, 但在不可见类别分割中精度不足, 森林与农田的混淆较为明显。SED 和 CAT-Seg 虽有所改进, 但仍存在边界模糊的问题。MSRINet 得益于全局-局部信息增强模块和旋转不变性特征优化模块, 有效提升了对复杂场景和不可见类别的适应能力, 其分割结果更接近真实标注。此外, MSRINet 在尺度较小的建筑区和农田上也展现了良好的预测效果。

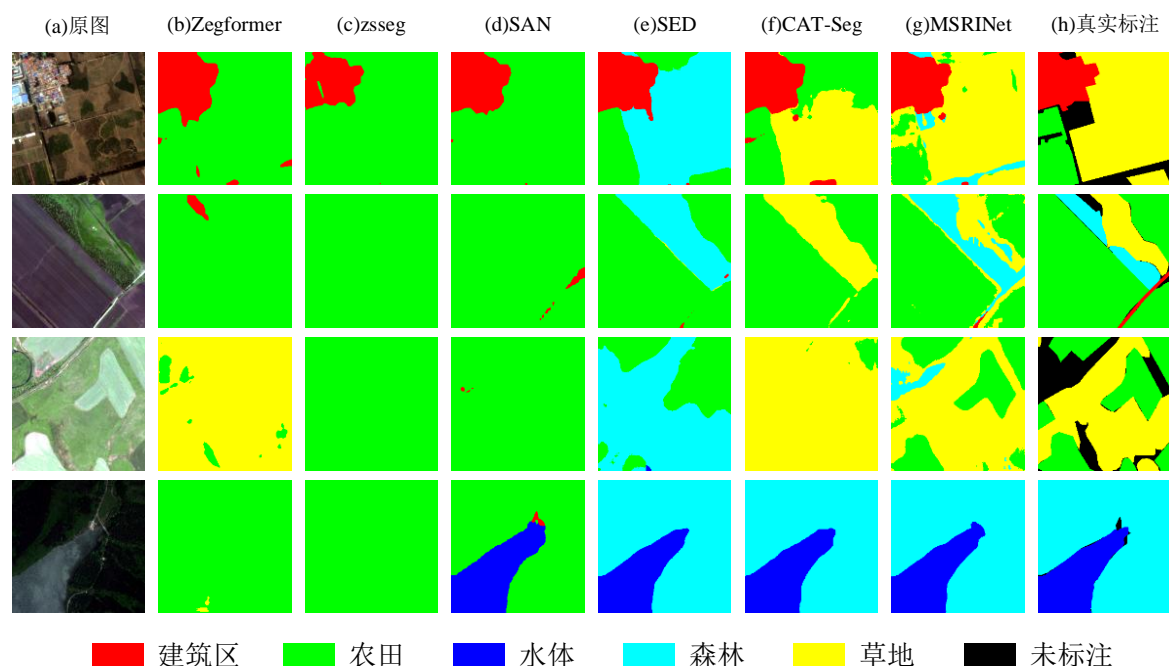


图 3.6 GID5 数据集上 MSRINet 与不同方法的预测结果

Fig.3.6 Prediction results of MSRINet and different methods on the GID5 dataset

表 3.3 提供了不同方法在 GID5 数据集上的详细预测指标，结果显示 MSRINet 在多个类别上均取得了优异表现，尤其是对于森林和草地的识别上，其 IoU 分别达到 88.94% 和 35.28%，在所有方法中位列第一。MSRINet 的总体 mIoU 和 mACC 分别达到 72.45% 和 84.40%，表明其在可见类别上的表现同样出色。相比之下，ZegFormer 在森林类别上的 IoU 仅为 1.24%，受限于两阶段分组的不足；zsseg、SAN 和 SED 在草地类别上的识别能力不足，且在可见类别上的精度也较低，前两者受阶段性设计限制，后者的层次编码优化在遥感场景下表现不足。CAT-Seg 在可见类别上表现良好，但在不可见类别森林和草地上，其 IoU 分别比 MSRINet 低 2.8% 和 4.57%。

表 3.3 GID5 数据集上 MSRINet 与不同方法的精度比较

Table 3.3 Comparison of the accuracies between MSRINet and different methods on the GID5 dataset

类别	Zegformer		zsseg		SAN		SED		CAT-Seg		MSRINet	
	I(%)	A(%)	I(%)	A(%)	I(%)	A(%)	I(%)	A(%)	I(%)	A(%)	I(%)	A(%)
建筑区	58.17	86.80	68.04	82.18	85.54	96.11	92.12	93.64	92.74	97.33	93.00	94.36
农田	31.94	87.32	32.03	99.05	45.46	98.39	78.13	84.20	81.35	91.21	68.64	71.37
水体	22.63	27.25	28.10	29.20	54.22	92.02	80.24	90.03	73.12	88.52	76.39	96.51
森林	1.24	2.05	14.15	14.25	34.14	34.48	78.25	96.77	75.20	83.22	88.94	96.00
草地	9.34	25.30	0.26	1.24	8.61	13.43	0.53	0.53	25.86	38.88	35.28	63.75
mIoU(%)	24.66		28.52		45.59		65.85		69.65		72.45	
mACC(%)	45.74		45.18		66.89		73.03		79.83		84.40	

注：表中 I 和 A 分别代表各类别的 IoU 和 ACC；双横线上的为可见类别；双横线下的为不可见类别；粗体代表最高精度值

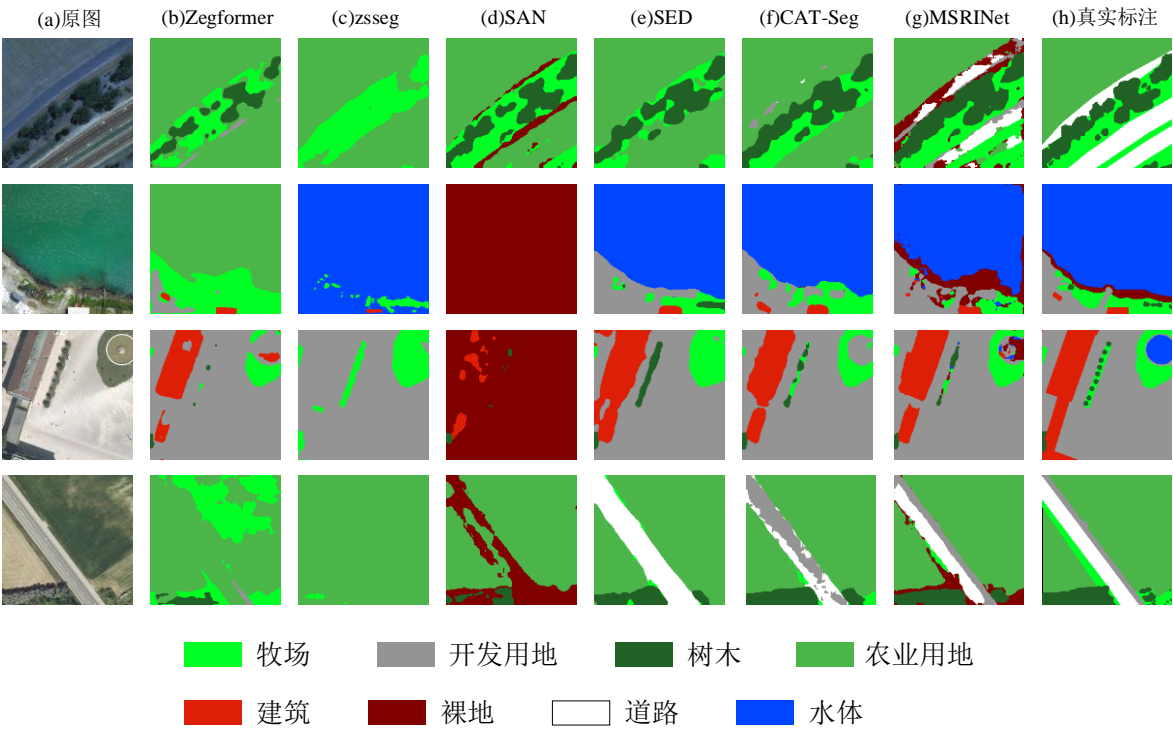


图 3.7 OpenEarthMap 数据集上 MSRINet 与不同方法的预测结果

Fig.3.7 Prediction results of MSRINet and different methods on the OpenEarthMap dataset

图 3.7 展示了不同方法在 OpenEarthMap 数据集上的预测结果，可视化结果表明，ZegFormer、Zsseg 和 SAN 等两阶段零样本语义分割方法在复杂地物分布下难以充分发挥 CLIP 的零样本能力，分割效果受限。相比之下，SED 和 CAT-Seg 通过计算语义匹配特征图能大致区分类别边界，但 SED 无法有效识别裸地类别，CAT-Seg 在道路分割上表现欠佳，前者受层次编码不足影响，后者对于语义匹配特征图的优化能力有限。整体而言，MSRINet 实现了最好的预测结果。

表 3.4 OpenEarthMap 数据集上 MSRINet 与不同方法的精度比较

Table 3.4 Comparison of the accuracies between MSRINet and different methods on the OpenEarthMap dataset

类别	Zegformer		zsseg		SAN		SED		CAT-Seg		MSRINet	
	I(%)	A(%)	I(%)	A(%)	I(%)	A(%)	I(%)	A(%)	I(%)	A(%)	I(%)	A(%)
农业用地	32.37	84.00	25.90	82.00	60.92	77.91	58.10	78.68	60.00	77.34	59.34	71.68
牧场	33.96	50.09	24.76	63.31	19.26	21.22	48.32	66.87	50.55	68.45	45.76	52.11
开发用地	35.16	65.77	18.50	27.17	20.42	24.57	41.96	72.12	42.14	72.81	44.59	59.92
建筑	58.79	76.51	10.02	10.16	40.72	44.11	69.34	82.18	71.27	84.89	66.68	85.64
树木	47.18	51.01	0.27	0.27	62.89	71.76	64.95	79.49	64.39	81.80	62.67	76.87
裸地	1.55	5.22	0.30	0.32	2.83	72.32	0.05	0.06	2.97	4.48	1.40	12.13
道路	0.77	0.99	14.81	25.69	3.36	3.42	15.45	17.25	7.05	7.23	16.37	18.63
水体	3.17	4.14	15.80	25.42	21.62	22.41	30.83	31.32	23.93	24.02	32.83	54.14
mIoU(%)	26.62		13.80		29.00		41.12		40.29		41.20	
mACC(%)	42.22		29.29		42.22		53.50		52.63		53.89	

注：表中 I 和 A 分别代表各类别的 IoU 和 ACC；双横线上的为可见类别；双横线下的为不可见类别；粗体代表最高精度值

与图 3.7 对应，表 3.4 展示了不同方法在 OpenEarthMap 数据集上的预测指标。在

OpenEarthMap 数据集上的实验结果表明, MSRINet 在遥感图像零样本语义分割任务中取得了较优异的性能。针对不可见类别(裸地、道路、水体), MSRINet 的 IoU 分别为 1.40%、16.37%和 32.83%相较于 Zegformer、zsseg 等方法表现出更高的分割精度, 尤其在水体类别上提升显著。同时, MSRINet 的 mIoU 和 mACC 分别达到 41.20%和 53.89%, 优于其他方法, 验证了其在不可见类别的泛化能力及整体分割效果的提升, 为遥感图像零样本语义分割研究提供了有效的解决方案。

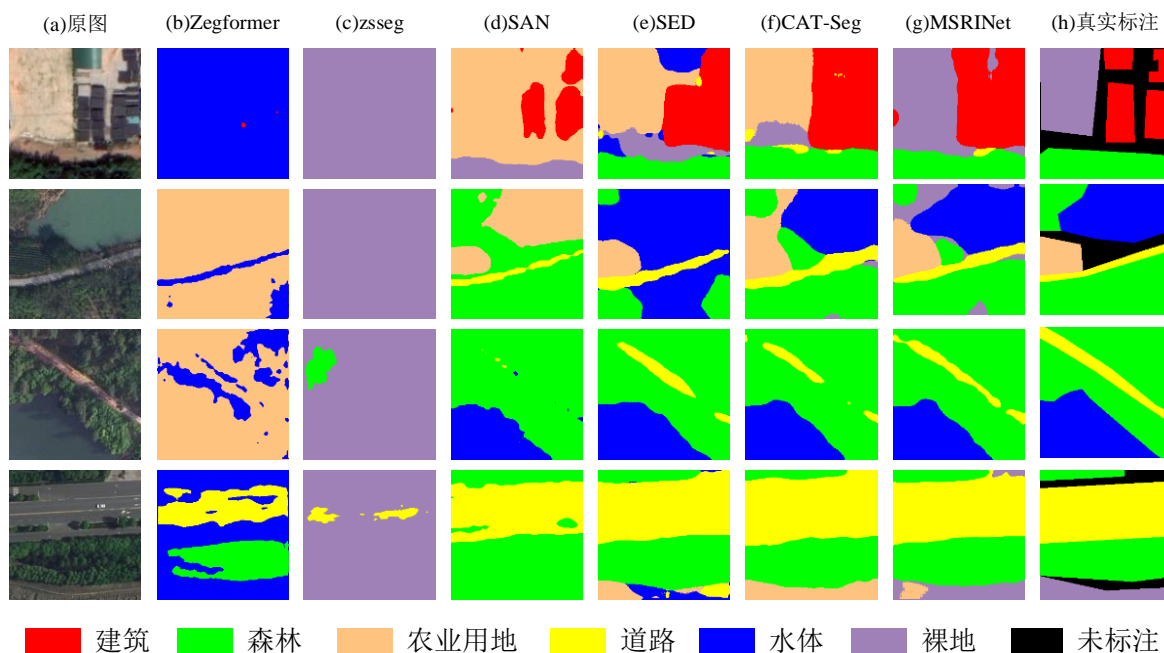


图 3.8 LoveDA 数据集上 MSRINet 与不同方法的预测结果

Fig.3.8 Prediction results of MSRINet and different methods on the LoveDA dataset

表 3.5 LoveDA 数据集上 MSRINet 与不同方法的精度比较

Table 3.5 Comparison of the accuracies between MSRINet and different methods on the LoveDA dataset

类别	Zegformer		zsseg		SAN		SED		CAT-Seg		MSRINet	
	I(%)	A(%)	I(%)	A(%)	I(%)	A(%)	I(%)	A(%)	I(%)	A(%)	I(%)	A(%)
建筑	8.35	8.40	25.86	32.71	65.16	88.99	74.75	90.55	70.57	92.34	72.68	92.54
道路	18.53	18.58	21.83	30.47	66.22	74.37	74.70	84.05	68.06	75.12	69.82	86.82
森林	8.34	8.42	12.65	34.20	33.11	70.28	59.80	67.67	63.75	78.36	64.36	76.90
农业用地	42.68	75.20	37.71	47.77	53.47	86.80	61.71	93.36	53.58	95.23	60.71	91.51
水体	22.63	50.40	0.01	0.01	15.58	16.15	38.70	57.80	11.99	12.01	21.30	21.42
裸地	0.08	0.15	15.47	49.89	15.68	16.74	1.50	1.51	35.80	48.16	40.02	68.63
mIoU(%)	16.77		18.92		41.54		51.86		50.63		54.81	
mACC(%)	26.86		32.51		58.89		65.82		66.87		72.97	

注: 表中 I 和 A 分别代表各类别的 IoU 和 ACC; 双横线上的为可见类别; 双横线下的为不可见类别; 粗体代表最高精度值

图 3.8 展示了不同方法在 LoveDA 数据集上的预测结果可视化。从结果来看, Zegformer 和 zsseg 表现出明显的类别混淆现象, 其中 Zegformer 倾向于将各类别错误预测为水体, 而 zsseg 则更倾向于预测为裸地。新类别语义的引入显著干扰了两者的预测准确性。相比之下, SAN、SED 和 CAT-Seg 在可见类别的预测上表现较好, 但在不可见类别

（如水体和裸地）的分割上仍显不足。而 MSRINet 在整体预测效果上更胜一筹，且对于道路这种尺度变化较大的类别仍能具有良好的提取效果，展现出更强的适应性和准确性。

表 3.5 进一步揭示了各方法在不可见类别（水体和裸地）分割性能上的显著差异。Zegformer 在水体和裸地的 IoU 分别仅为 22.63%和 0.08%，反映出其对未见类别的分割能力较弱；zsseg 在水体上的 IoU 甚至低至 0.01%，表现尤为逊色。相比之下，SAN、SED、CAT-Seg 和 MSRINet 在这些类别上的表现有所改善，其中 MSRINet 尤为突出，其在裸地的 IoU 达到 40.02%，水体为 21.30%，显示出较强的泛化能力。MSRINet 的整体 mIoU 和 mACC 分别达到 54.81%和 72.97%，在所有方法中位居首位，凸显其优越的性能。

3.3.6 消融实验分析

本节通过消融实验验证了 MSRINet 各模块在零样本语义分割中的关键作用，实验结果如表 3.6 所示。消融分析采用逐步添加模块的方式，旨在评估每个组件对模型性能的独立贡献及其协同效应。以仅利用 CLIP 视觉-文本嵌入生成语义匹配特征图并输入分割头的方案为基线，其在 GID5、OpenEarthMap 和 LoveDA 数据集上的 mIoU 分别为 61.49%、35.15%和 42.46%，在复杂遥感场景中泛化能力有限。

引入全局-局部信息增强模块（GLIEM）后，通过融合 CLIP 的全局语义特征与 ResNet 的局部细节特征，显著提升了多尺度特征表达能力，例如在 LoveDA 数据集上 mIoU 提升至 49.74%，表明多尺度融合有效改善了模型对不同尺度地物的适应性。接着加入旋转不变性特征优化模块（RIFOM），通过多角度旋转和空间注意力机制优化特征鲁棒性，使模型在 GID5 上的 mIoU 增至 64.47%。进一步引入类别关系建模模块（CRMM）后，利用自注意力机制挖掘类别间的语义关联，优化了语义表达能力，在 OpenEarthMap 上 mIoU 提升至 39.12%。

表 3.6 MSRINet 的消融实验
Table 3.6 The abalition experiment of MSRINet

方法	GID5		OpenEarthMap		LoveDA	
	mIoU(%)	mACC(%)	mIoU(%)	mACC(%)	mIoU(%)	mACC(%)
基线方法	61.49	76.56	35.15	48.16	42.46	58.15
基线方法+GLIEM	63.45	78.26	35.49	47.49	49.74	64.41
基线方法+RIFOM	64.47	80.55	37.57	49.83	46.74	66.41
基线方法+CRMM	63.89	77.57	39.12	55.70	48.18	59.98
基线方法+GLIEM+RIFOM	69.55	88.60	40.85	47.55	50.50	62.15
基线方法+GLIEM+CRMM	68.11	82.13	41.15	51.42	53.16	69.16
基线方法+GLIEM+RIFOM+CRMM	72.45	84.40	41.20	53.89	54.81	72.97

注：粗体代表最高精度值

当 GLIEM、RIFOM 和 CRMM 三个模块协同作用时，MSRINet 性能达到最佳状态，在 LoveDA 上的 mIoU 和 mACC 分别提升至 54.81%和 72.97%，在 GID5 上 mIoU 增至 72.45%，在 OpenEarthMap 上也显著优于基线。实验结果充分证明，GLIEM 增强了特征的

多尺度表达，RIFOM 提升了模型对复杂场景的鲁棒性，CRMM 优化了类别语义的精细性，三者组合显著提升了 MSRINet 在零样本语义分割任务中的整体性能，为遥感图像分割提供了高效且可靠的解决方案。

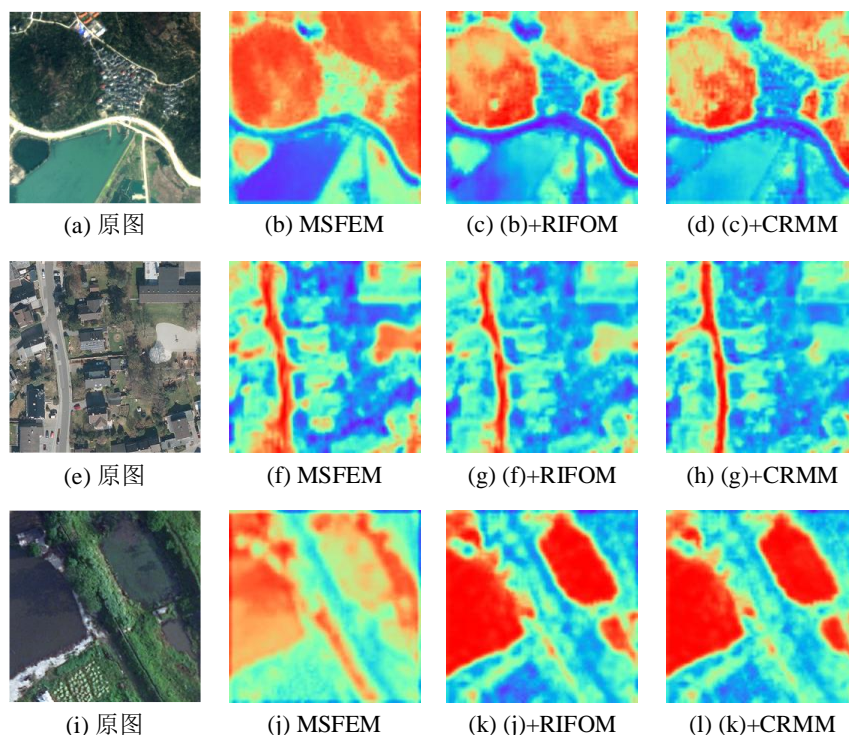


图 3.9 模型中间过程特征图可视化

Fig.3.9 Visualization of the Feature Maps in the Intermediate Process of the Model

为了更直观地验证全局-局部信息增强模块（GLIEM）、旋转不变性特征优化模块（RIFOM）以及类别关系建模模块（CRMM）在 MSRINet 中的有效性，本节对各模块的中间特征图进行了可视化分析。对于 GLIEM 模块，选择了语义匹配特征图进行可视化；对于 RIFOM 模块，选择了 RIFOM 模块的输出 F_{RIFO} 进行可视化；对于 CRMM 模块，选择了 CRMM 模块的输出 F_{CRM} 进行可视化。为了评估模型对未见类别的零样本分割能力，每张输入图像均针对某一不可见类别生成对应的特征图，可视化结果如图 3.9 所示。

对于输入图像 (a) 中的“森林”类别，GLIEM 模块通过融合 CLIP 视觉编码器的全局语义特征与 ResNet 的局部细节特征，能够初步提取森林区域的空间分布。结果表明，该模块不仅成功捕获了大尺度连片森林区域，还能识别建筑群间的小尺度森林区域。然而，此时特征图对森林轮廓的刻画能力较弱，边界细节模糊。随后，经过 RIFOM 模块的多角度旋转操作和空间注意力机制处理，显著优化了森林类别的轮廓表征，增强了特征的旋转不变性，使边界更加清晰。最后，CRMM 模块通过自注意力机制挖掘类别间的语义关系，进一步减少了类别混淆现象。例如，在图 (d) 的左下角，水体类别被成功从森林类别中分离出来，表明 CRMM 有效提升了类别区分能力。

对于图像 (e) 中的“道路”类别，其提取过程具有一定挑战性。GLIEM 模块能够初步定位道路区域的空间分布，但生成的语义匹配特征图在道路连续性和轮廓精确性方面存

在不足，表现为道路特征的不连续和边界模糊。经过 RIFOM 模块处理后，引入旋转不变特征，显著改善了道路区域的连续性，并增强了特征的空间一致性。随后，CRMM 模块通过优化类别间的语义表征，进一步细化了道路特征，减少了与相似类别（如“裸地”）的混淆现象。最终输出展示了更精确的道路分割结果，验证了 RIFOM 和 CRMM 在提升特征表征质量方面的协同作用。

对于输入图像 (i) 中的“森林”类别，由于森林与水体在颜色上的相似性，仅依赖 GLIEM 模块提取的特征图存在一定程度的类别混淆，部分森林区域被错误识别为水体。经过 RIFOM 模块处理后，通过旋转不变性特征的引入，初步区分了森林和水体两种类别，缓解了混淆问题。随后，CRMM 模块进一步利用类别间的语义关系对特征进行优化，展现了更清晰的类别边界和更高的分割精度，森林区域的提取效果得到显著提升。

3.3.7 多尺度特征层选取分析

为探究全局-局部信息增强模块（GLIEM）中 CLIP 中间层选择对遥感图像零样本语义分割性能的影响，本节设计了对照实验，分别评估了第 1、2、3 层，第 1、3、5 层，以及第 1、4、7 层三种组合的效果，实验结果如表 3.7 所示。实验在 GID5、OpenEarthMap 和 LoveDA 三个数据集上进行，以 mIoU 和 mACC 为主要指标，旨在分析不同层级特征组合对模型分割能力的作用。

表 3.7 CLIP 中间层选择对不同数据集分割性能的影响

Table 3.7 Impact of Selected CLIP Intermediate Layers on the Segmentation Performance of Different Datasets

选取的 CLIP 中间层	GID5		OpenEarthMap		LoveDA	
	mIoU(%)	mACC(%)	mIoU(%)	mACC(%)	mIoU(%)	mACC(%)
1, 2, 3	71.49	81.49	40.48	49.84	53.45	68.15
1, 3, 5	72.45	84.40	41.20	53.89	54.81	72.97
1, 4, 7	70.65	76.17	38.16	45.16	52.16	65.16

注：粗体代表最高精度值

首先，第 1、2、3 层组合聚焦于 CLIP 的前三层浅层特征，擅长提取低级视觉信息（如边缘、纹理和局部结构），能有效捕捉遥感图像中的细节特征。然而，由于层间跨度较小，语义抽象能力有限，在需要较强语义理解的数据集（如 OpenEarthMap）上表现较差。其次，第 1、3、5 层组合在所有数据集上均展现出最佳性能。该组合跨度适中，涵盖从浅层到中层的特征，既保留了丰富的空间细节，又增强了语义表达能力。通过特征金字塔网络（FPN）的跨尺度融合，该组合实现了局部细节与全局语义的有效平衡，显著提升了模型的适应性和鲁棒性。最后，第 1、4、7 层组合的性能相对较弱。深层特征（第 7 层）因过于抽象，丢失了分割任务所需的分辨率和细节信息，而层间跨度过大导致浅层与深层特征难以协调融合，削弱了模型对小尺度目标的捕捉能力，尤其在 OpenEarthMap 这样细节丰富的遥感数据集上表现不佳。

基于以上分析，第 1、3、5 层组合因兼顾细节保留与语义理解，成为 GLIEM 模块的

最优选择，为遥感图像零样本语义分割提供了高效的特征提取策略。

3.3.8 旋转角度对 RIFOM 的影响分析

为探究旋转角度对旋转不变性特征优化模块（RIFOM）性能的影响，本节设计了多组对照实验，通过设置不同数量的旋转角度（2 个、3 个、4 个）验证 RIFOM 的有效性，实验结果如表 3.8 所示。实验在 GID5、OpenEarthMap 和 LoveDA 三个遥感数据集上进行，以 mIoU 和 mACC 作为评价指标。

表 3.8 RIFOM 中旋转度数对不同数据集分割性能的影响

Table 3.8 The Influence of the Rotation Degrees in RIFOM on the Segmentation Performance of Different Datasets

旋转度数	GID5		OpenEarthMap		LoveDA	
	mIoU(%)	mACC(%)	mIoU(%)	mACC(%)	mIoU(%)	mACC(%)
0°, 90°	68.20	82.15	39.04	49.79	52.98	69.50
0°, 90°, 180°	71.56	82.13	41.38	52.19	53.30	74.16
0°, 90°, 180°, 270°	72.45	84.40	41.20	53.89	54.81	72.97

注：粗体代表最高精度值

实验结果表明，旋转角度数量的增加对 RIFOM 性能提升具有显著的正向作用。在 GID5 数据集上，mIoU 从两个角度的 68.20% 提升至三个角度的 71.56%，并在四个角度时达到峰值 72.45%，mACC 则从 82.15% 稳步增至 84.40%，显示更多旋转角度显著增强了模型对复杂地物的分割精度和分类能力。在 OpenEarthMap 数据集上，mIoU 从两个角度的 39.04% 逐步提高到三个角度的 41.38%，最终在四个角度时稳定于 41.20%，而 mACC 从 49.79% 持续上升至 53.89%，表明旋转角度的增加有效提升了模型对不可见类别的适应性。同样，在 LoveDA 数据集上，mIoU 从两个角度的 52.98% 稳步提升至四个角度的 54.81%，mACC 从 69.50% 增至 72.97%，进一步验证了这一趋势，尤其在多视角场景下的鲁棒性更强。

RIFOM 通过多角度旋转操作生成等变特征，并结合空间注意力机制提取旋转不变特征，角度数量的增加丰富了特征表示的多样性，从而提升了模型对遥感图像中视角变化的适应能力。综上所述，旋转角度越多，RIFOM 在提取旋转不变特征方面的能力越强，在所有测试数据集上均实现了更高的 mIoU 和 mACC，四个角度被证明是最优选择。

3.4 本章小结（Summary）

本章通过在 GID5、OpenEarthMap 和 LoveDA 三个遥感数据集上的实验，全面验证了 MSRINet 在零样本语义分割任务中的优越性能。MSRINet 通过全局-局部信息增强模块（GLIEM）、旋转不变性特征优化模块（RIFOM）和类别关系建模模块（CRMM）的协同作用，有效解决了遥感图像多尺度目标、旋转不变性和复杂语义关系的挑战。实验结果表明，MSRINet 在不可见类别的分割上表现出色，例如在 GID5 数据集上森林和草地的 IoU 分别达到 88.94% 和 35.28%，整体 mIoU 和 mACC 分别达到 72.45% 和 84.40%，显著优

于 ZegFormer、zsseg、SAN、SED 和 CAT-Seg 等方法。消融实验进一步证明了各模块的有效性，为遥感图像零样本语义分割提供了高效、鲁棒的解决方案，展现了其在环境监测、城市规划等领域的应用潜力。

4 基于知识图谱和信息瓶颈理论的零样本语义分割方法

4 Knowledge Graph and Information Bottleneck-based Zero-Shot Semantic Segmentation

4.1 引言 (Foreword)

零样本语义分割通过利用外部知识挖掘类别间的语义关系或跨模态信息，将可见类别的知识迁移至不可见类别，在诸多领域展现出重要应用潜力，然而其实现面临诸多技术挑战。基于 CLIP 模型在零样本任务中表现出较强的跨模态对齐能力，通过在大规模自然图像和文本数据上的预训练，能够有效关联视觉特征与语义标签。Zhou 等人设计单阶段 ZegCLIP 框架，提出基于文本-块嵌入相似度计算的像素级预测范式，通过特征空间校准、语义原型解耦及上下文感知增强三重机制优化 CLIP 的跨粒度迁移能力^[13]。He 等人提出基于基元生成与跨空间对齐的通用框架，设计语义引导的基元组合特征生成器，并实施语义相关和无关视觉特征解耦与跨模态对齐约束，实现语义-视觉空间的深度耦合^[62]。然而，CLIP 的训练数据主要来源于自然场景，其视觉特征和类别体系与遥感图像存在明显差异。具体而言，遥感图像中的地物类别具有领域特异性，依赖专业术语和特定分类体系，且其语义内涵与地物的空间分布及内在属性紧密相关，而这些特性难以通过 CLIP 基于通用的类别标签充分表征。如图 4.1 所示，仅使用通用类别标签进行预测，开发用地类别和建筑物类别出现了大量混淆。此外，遥感图像的传感器类型、光谱特性及成像条件与自然图像差异较大，导致相同地物类别在两种场景下的视觉特征可能存在显著不同。这些因素使得 CLIP 在遥感图像零样本语义分割中难以准确捕捉类别语义或生成鲁棒的视觉特征表示，尤其在处理不可见类别及复杂地物分布时表现更为明显。

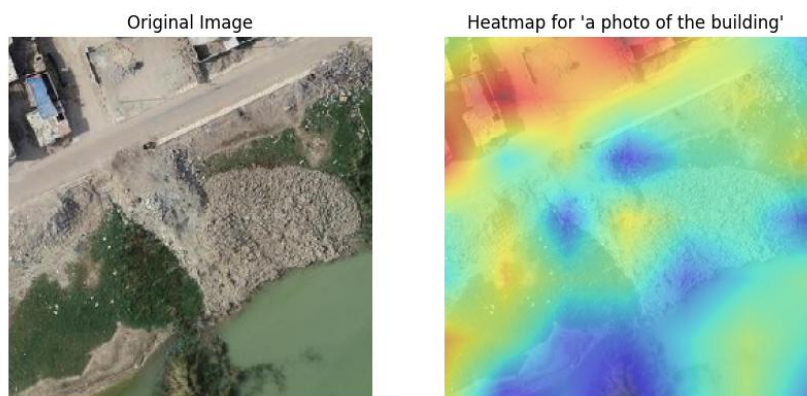


图 4.1 仅使用 CLIP 通用类别标签预测的原图像和热力图

Fig.4.1 The original image and the heatmap predicted using only the CLIP general category labels

针对上述挑战，知识图谱作为一种结构化的知识表示形式，为解决遥感图像零样本语义分割的语义建模难题提供了解决方案。知识图谱通过节点和边的形式描述地物类别

及其关系，能够表征类别间的语义关联。例如，它可以揭示“河流”与“湖泊”在属性上的相似性，或区分“建筑物”与自然地物在语义和物理特征上的差异，从而为文本嵌入提供丰富的上下文信息。同时，知识图谱通过编码空间关系和节点属性，能够辅助优化视觉特征表示，弥补传统方法在表征地物空间分布及多维属性方面的不足。然而，将知识图谱的文本语义特征与遥感图像的视觉特征进行有效融合仍面临难题。由于两者分属不同模态，如何增强模态间的相关性并减少冗余信息成为关键技术挑战。

为解决模态融合中的冗余问题，信息瓶颈理论（Information Bottleneck, IB）被引入作为优化策略。信息瓶颈理论通过压缩文本语义特征中的无关信息并保留与视觉特征高度相关的内容，能够提升两种模态之间的相关性，同时降低模态间的冗余。具体而言，该理论可筛选知识图谱中与视觉模态直接相关的语义属性和空间关系，剔除无关信息，从而增强跨模态表示的精准性和一致性。基于此，提出了一种基于知识图谱和信息瓶颈理论的零样本语义分割方法（Knowledge Graph and Information Bottleneck Network, KGIBNet）。该方法设计了属性知识增强模块（Attribute Knowledge Enhancement Module, AKEM），利用知识图谱的语义信息提升文本嵌入对不可见类别的分辨能力；同时，借助信息瓶颈理论优化文本特征与视觉特征的融合过程，进一步提高跨模态表示的质量。

4.2 知识图谱与信息瓶颈网络 KGIBNet (Knowledge graph and information bottleneck network)

4.2.1 KGIBNet 整体架构

本章提出了一种基于知识图谱和信息瓶颈理论的零样本语义分割网络（Knowledge Graph and Information Bottleneck Network, KGIBNet），整个网络的示意图如 4.2 所示。该网络首先通过属性知识增强模块（Attribute Knowledge Enhancement Module, AKEM）提取知识图谱中的属性信息，生成知识嵌入，并将其与 CLIP 的文本嵌入融合，得到增强的文本语义嵌入，以提升对不可见地物类别的语义分辨能力。这些文本语义嵌入不仅包含类别名称的基本语义信息，还整合了知识图谱中提取的领域特异性知识。其次，利用文本语义嵌入与 CLIP 视觉编码器提取的视觉嵌入，通过余弦相似度计算语义匹配特征图。该特征图表征了图像视觉特征与文本语义信息之间的匹配关系，可视为粗略的语义掩码视觉表示。随后，基于知识图谱的图结构，采用图卷积网络（Graph Convolutional Network, GCN）处理文本语义嵌入，引入知识图谱的空间分布信息以生成空间分布嵌入；同时，通过多头注意力机制进一步提取文本语义嵌入中与类别相关的属性信息，生成细化属性嵌入。接着，设计了基于信息瓶颈的知识融合模块（Knowledge Fusion Based on Information Bottleneck, KFBIB），将语义匹配特征图与来自文本模态的空间分布嵌入和细化属性嵌入进行融合，通过筛选与语义匹配特征图相关的信息并滤除冗余信息，生成适用于遥感图像的特征表示。该跨模态融合过程旨在增强文本语义信息与视觉特征之间的

交互，从而提升模型在复杂遥感场景中对不可见地物类别的零样本语义分割性能。

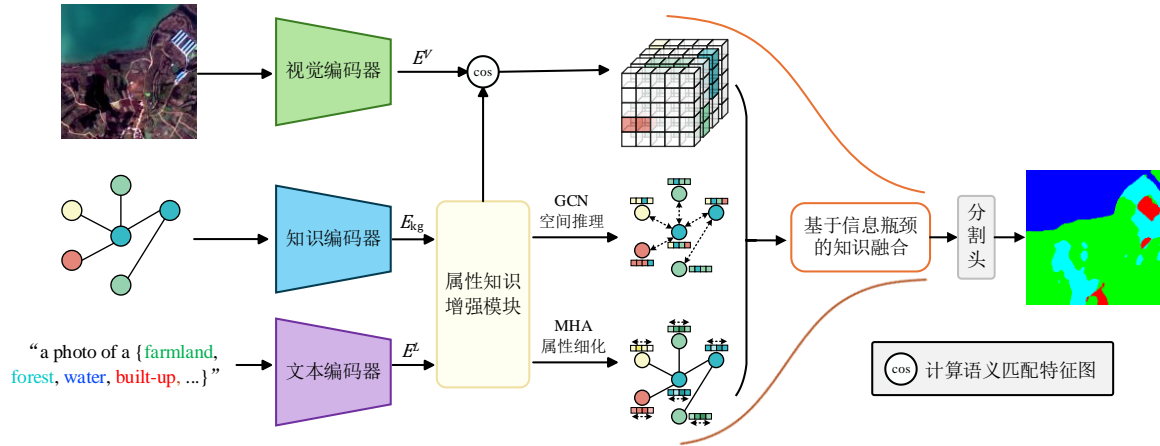


图 4.2 KGIBNet 整体框架

Fig.4.2 The overall architecture of KGIBNet

4.2.2 知识图谱构建

遥感地学知识图谱的构建旨在为零样本遥感图像语义分割任务提供结构化的知识表示，通过系统化地描述地物类别、语义关系及空间分布，增强模型对不可见类别的泛化能力。构建过程融合 RSKG 和 ConceptNet 作为主要数据来源，并结合专家知识，针对特定类别需求精心设计了 90 条三元组，为跨领域知识迁移奠定基础。其中空间分布关系共 34 条三元组，涵盖“临近”（如“建筑物临近道路”）、“包围”（如“农业用地包围水体”）等关系，捕捉地物间的空间拓扑特性；而属性关系共 56 条三元组，包括“类型”（如“建筑物是人工地物”）、“颜色”（如“森林为绿色”）等属性，强化地物的语义描述。这些三元组通过节点（地物实体）和边（关系）构建层次化的语义网络，编码多维属性和空间分布规律，支持模型推断未标记地物的语义和空间特性，从而显著提升零样本任务中的识别性能。

4.2.3 属性知识增强模块

遥感场景中的地物类别名称具有显著的领域特异性，仅依赖类别名称难以充分表征与视觉特征相匹配的深层语义信息。为应对这一挑战，引入知识图谱中的属性信息作为外部语义信息的来源，以结构化的形式全面地描述类别名称所蕴含的丰富语义内涵。对于某个类别 c ，在知识图谱中检索出以该类别为头实体的全部属性三元组 $T_c = \{(h_i, r_i, t_i) | i \in [1, N_t]\}$ ，其中 h_i, r_i, t_i, N_t 分别表示头实体（即类别 c 的名称）、关系实体、尾实体以及以该类别为头实体的三元组数量。

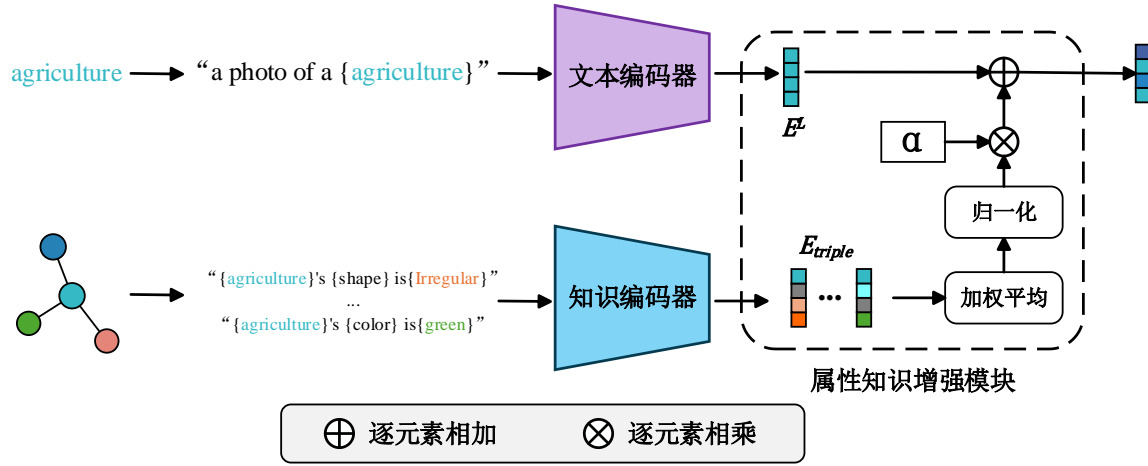


图 4.3

属性知识增强模块

Fig.4.3 Attribute Knowledge Enhancement Module

为了将知识图谱中的语义信息转化为可供模型处理的嵌入，采用提示模板策略将每个三元组中的属性关系转化为自然语言句子。具体而言，将属性关系使用模板为“{头实体}'s {关系} is {尾实体}”转化为句子。通过以下两步生成类别 c 的知识嵌入 E_{kg} 。

首先，利用 CLIP 文本编码器 $\Phi^L(\cdot)$ 对每个生成的句子进行编码，生成三元组嵌入 E_{triple_i} ：

$$E_{triple_i} = \Phi^L(\text{prompt}(h_i, r_i, t_i)) \quad (4.1)$$

其中， $\text{prompt}(h_i, r_i, t_i)$ 表示属性关系模板。

其次，对类别 c 的三元组嵌入 $E_{triple_1}, \dots, E_{triple_{N_t}}$ 进行加权平均并归一化，得到属性知识嵌入 E_{kg} ，其计算过程如下：

$$E_{kg} = \frac{\sum_{i=1}^{N_t} w_c \cdot E_{triple_i}}{\|\sum_{i=1}^{N_t} w_c \cdot E_{triple_i}\|} \quad (4.2)$$

其中， w_c 是通过自学习机制优化的类别权重，用于动态调整同一类别不同属性信息的重要程度。

为了进一步融合属性知识嵌入与类别名称的基础语义信息，同样利用提示模板策略生成类别文本嵌入 E^L ，即：

$$E^L = \Phi^L(\text{prompt}(c)) \quad (4.3)$$

其中， $\text{prompt}(c)$ 采用默认模板“a photo of the {类别名称}”以适配 CLIP 的预训练特性。

最终，通过将属性知识嵌入 E_{kg} 与类别文本嵌入 E^L 进行加权融合，得到文本语义嵌入 E_{kg}^L ：

$$E_{kg}^L = E_{kg} + \alpha \cdot E^L \quad (4.4)$$

其中， α 是自学习的加权系数，用于平衡知识图谱信息与类别名称信息的重要性。

通过上述属性知识增强模块，获得的 E_{kg}^L 不仅包含类别名称的基础语义信息，还整

合了从知识图谱中提取的领域特异性知识，从而增强了模型对遥感图像中类别的理解与区分能力。

与第 3 章相同，通过计算语义匹配特征图辅助分割任务。给定一幅图像 I 和一组类别 c ，首先通过 CLIP 图像编码器 $\Phi^V(\cdot)$ 提取密集的图片嵌入 $E^V = \Phi^V(I) \in \mathbb{R}^{H \times W \times d}$ ，其中 H 和 W 为图像的高度和宽度， d 为图像嵌入的维度。接下来，利用图像嵌入 E^V 与文本语义嵌入 E_{kg}^L 进行余弦相似度计算，生成语义匹配特征图 $F_{\cos} \in \mathbb{R}^{H \times W \times N_c}$ ：

$$F_{\cos} = \frac{E^V \cdot E_{kg}^L}{\|E^V\| \|E_{kg}^L\|} \quad (4.5)$$

该语义匹配特征图能够有效捕捉图像局部视觉特征与类别语义嵌入之间的匹配度，为后续零样本语义分割提供精确依据。为了进一步增强语义匹配特征图在高维空间的处理能力以适应精细分割任务，在类别尺度上引入卷积层对 F_{\cos} 进行处理，生成初始的语义匹配特征体 $F_{\cos} \in \mathbb{R}^{H \times W \times N_c \times d}$ ，其中 d 是嵌入维度。

4.2.4 基于信息瓶颈的知识融合

为进一步优化语义匹配特征体，使其更好地捕捉遥感场景中不同类别间的空间分布关系和属性特征，引入知识图谱中的空间关系作为补充信息源。然而，直接利用知识图谱进行优化面临两方面的难点：一是来自视觉模态的 F_{\cos} 与来自文本模态 E_{kg}^L 的之间的固有差异性，二是两模态之间可能存在无关的冗余信息。为解决这些问题，引入信息瓶颈理论（Information Bottleneck, IB），旨在筛选与语义匹配特征体高度相关的文本语义信息并滤除冗余信息，从而生成适用于遥感图像的多模态特征表示。

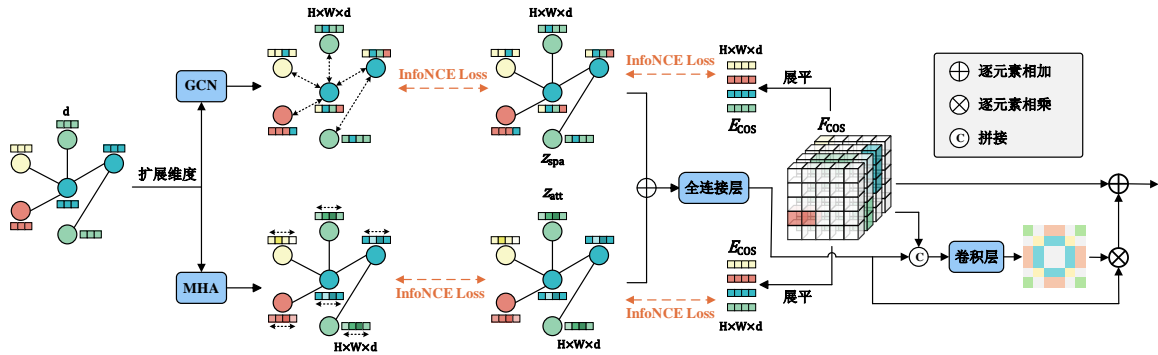


图 4.4 基于信息瓶颈的知识融合

Fig.4.4 Knowledge Fusion Based on Information Bottleneck

基于信息瓶颈理论的优化目标是通过最大化知识增强的文本语义嵌入与视觉特征之间的互信息，同时最小化文本语义嵌入与视觉特征之间的互信息，从而优化类别语义。具体而言，优化过程从空间和属性两个维度对知识图谱信息进行建模。采用图卷积网络（Graph Convolutional Network, GCN）和多头自注意力机制（Multi-Head Attention, MHA）分别生成空间分布嵌入和细化属性嵌入。在空间维度上，GCN 基于知识图谱的空间分布

图结构，通过推理文本语义嵌入迭代传播类别间的邻接关系，从而捕捉遥感图像中地物类别的空间分布特征。在属性维度上，MHA 进一步优化文本语义嵌入中的属性信息，生成细粒度的属性表征，有效表征遥感图像中地物类别的多维特征，例如形状、光谱和纹理等。

首先，对于空间维度，将优化后的空间嵌入 z_{spa} 、经过 GCN 推理后的知识增强嵌入 $GCN(E_{kg}^L)$ 、以及语义匹配特征嵌入 E_{COS} （即将语义匹配特征体 F_{COS} 沿着空间维度展平）之间的关系通过以下损失函数建模：

$$\mathcal{L}_{spa} = -I(z_{spa}; E_{COS}) + \beta I(z_{spa}; GCN(E_{kg}^L)) \quad (4.6)$$

其中， $I(z_{spa}; E_{COS})$ 表示空间嵌入与语义匹配特征嵌入的互信息，采用 InfoNCE 损失计算，旨在最大化两者之间的相关性； $I(z_{spa}; GCN(E_{kg}^L))$ 表示空间嵌入与经过 GCN 推理后的文本语义嵌入之间的互信息，同样通过 InfoNCE 损失计算，旨在最小化冗余信息； β 是平衡系数，用于调节相关性和冗余之间的权衡。对于属性维度，优化过程类似，其损失函数为：

$$\mathcal{L}_{att} = -I(z_{att}; E_{COS}) + \beta I(z_{att}; MHA(E_{kg}^L)) \quad (4.7)$$

其中， z_{att} 表示优化后的属性嵌入。于是，本章的最终损失为：

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda \cdot (\mathcal{L}_{spa} + \mathcal{L}_{att}) \quad (4.8)$$

其中， \mathcal{L}_{CE} 为交叉熵损失， λ 为超参数。

在优化完成后，空间嵌入 z_{spa} 和属性嵌入 z_{att} 通过拼接的方式进行融合，并经过一个全连接层生成类别知识语义表征 F_{IB} 。其计算过程如下：

$$F_{IB} = FC(\text{cat}(z_{spa}, z_{att})) \quad (4.9)$$

其中，FC 表示全连接层序列，包含非线性激活函数。

为了进一步优化初始语义匹配特征体 F_{COS} ，通过注意力机制将类别语义特征图 F_{IB} 与初始语义匹配特征体 F_{COS} 进行融合。具体而言，类别语义特征图首先通过线性映射层映射到与特征图相同的通道维度，并扩展到相同的空间维度；随后，通过一个注意力网络生成空间和属性相关的权重，用于增强特征图的表达能力。最终，优化后的特征图 F_{KGIB} 通过以下方式生成：

$$F_{KGIB} = F_{COS} + \text{Attention}(\text{cat}(F_{COS}, F_{IB})) \quad (4.10)$$

其中，Attention 表示注意力机制，由卷积层和激活函数组成，用于动态调整特征图的权重。

通过上述优化过程，优化后的语义匹配特征体 $F_{KGIB} \in \mathbb{R}^{H \times W \times N_c \times d}$ 不仅保留了初始视觉特征与语义嵌入的匹配信息，还通过知识图谱的空间和属性信息增强了类别间的语义关联，同时利用信息瓶颈理论有效减少了冗余信息的影响，从而显著提升了遥感场景下

零样本语义分割的性能。最后，使用一个简单的二维卷积将 d 维映射到 1 维并上采样到原图分辨率，生成最终的分割预测结果。KGIBNet 的流程在算法 2 中进行了详细阐述。

算法 2: 基于知识图谱和信息瓶颈理论的零样本语义分割

输入:

I : 遥感图像

KG : 包含类别三元组的知识图谱 (覆盖可见和不可见类别)

C_s : 可见类别

C_u : 不可见类别

GT_s : 可见类别的像素级标注 (训练时使用, 仅对 C_s 对应的类别)

输出:

S : 输入图像的语义分割掩码 (包含不可见类别)

算法:

1. 训练阶段 (仅针对可见类别):
 - 1) 输入训练图像 I_s 、 KG 和对应的可见类别标注 GT_s
 - 2) 属性知识增强模块 (AKEM):
 - a) 从 KG 中提取与 C_s 相关的三元组, 转换为自然语言句子
 - b) 使用 CLIP 文本编码器生成属性知识嵌入 E_{kg}
 - c) 加权结合 C_s 的类别文本嵌入 E_c , 生成 E_{kg}^L
 - 3) 使用 CLIP 视觉编码器从 I_s 提取视觉特征 E^V
 - 4) 使用 E^V 和 E_{kg}^L 进行余弦相似度计算并卷积, 从而获得语义匹配特征体 F_{cos}
 - 5) 基于信息瓶颈的知识融合 (KFBIB):
 - a) 使用 GCN 和 KG 的邻接矩阵建模空间关系
 - b) 使用多头自注意力细化属性信息
 - c) 使用信息瓶颈优化空间嵌入 z_{spa} 和属性嵌入 z_{att}
 - d) 将 z_{spa} 和 z_{att} 进行拼接并和 F_{cos} 进行注意力加权融合
 - 6) 使用 GT_s 监督训练, 优化模型参数
2. 测试阶段 (针对所有类别):
 - 1) 输入测试图像 I_u 、可见类别 C_s 、不可见类别 C_u 和 KG
 - 2) 经过 KGIBNet 推理生成分割掩码 S
3. 返回 S

4.3 实验结果与分析 (Experimental results and analysis)

4.3.1 对比实验分析

本章实验环境及相关参数设置同第三章实验环境。为了验证该模型的有效性, 本文

在 GID5、OpenEarthMap 和 LoveDA 数据集上与一些相关工作进行了比较。其中包括单阶段的方法 ZegCLIP^[13] (CVPR2023), 两阶段方法 PADing^[62](CVPR2023)、SAN^[59](CVPR2023)和计算语义匹配特征体的 SED^[60](CVPR2024)、Cat-Seg^[61](CVPR2024)方法。

实验结果如图 4.5 所示, 不同模型在零样本语义分割任务中的表现呈现出显著差异。ZegCLIP 和 PADing 在整体预测性能上表现较为优异, 得益于其对 CLIP 预训练特征的有效利用。然而, ZegCLIP 在处理未见类别(如草地)时表现出明显的预测偏倚, 倾向于将边界区域过度分配给草地, 导致类别边界划分的精确性不足。PADing 虽通过生成式方法缓解了未见类别的数据缺失问题, 但在复杂场景中仍受限于语义-视觉空间的领域差距, 边界分割的细腻度有待提升。SAN 模型在不可见类别的识别能力上表现较弱, 尤其在农田与草地、农田与森林等语义相近的类别间, 出现了严重的混淆现象。SED 和 CAT-Seg 模型在可见类别的分割任务中展现出较强的性能, 然而 SED 对草地的预测准确率偏低, 可能是层次编码器在全局上下文建模上的局限性导致未见类别的语义表征不足, CAT-Seg 则在森林边界的精细化分割上表现不足。相比之下, KGIBNet 通过引入知识图谱驱动的语义校正机制, 有效解决了森林、草地等不可见类别与农田之间的语义混淆问题, 最终实现了最优的零样本语义分割性能。

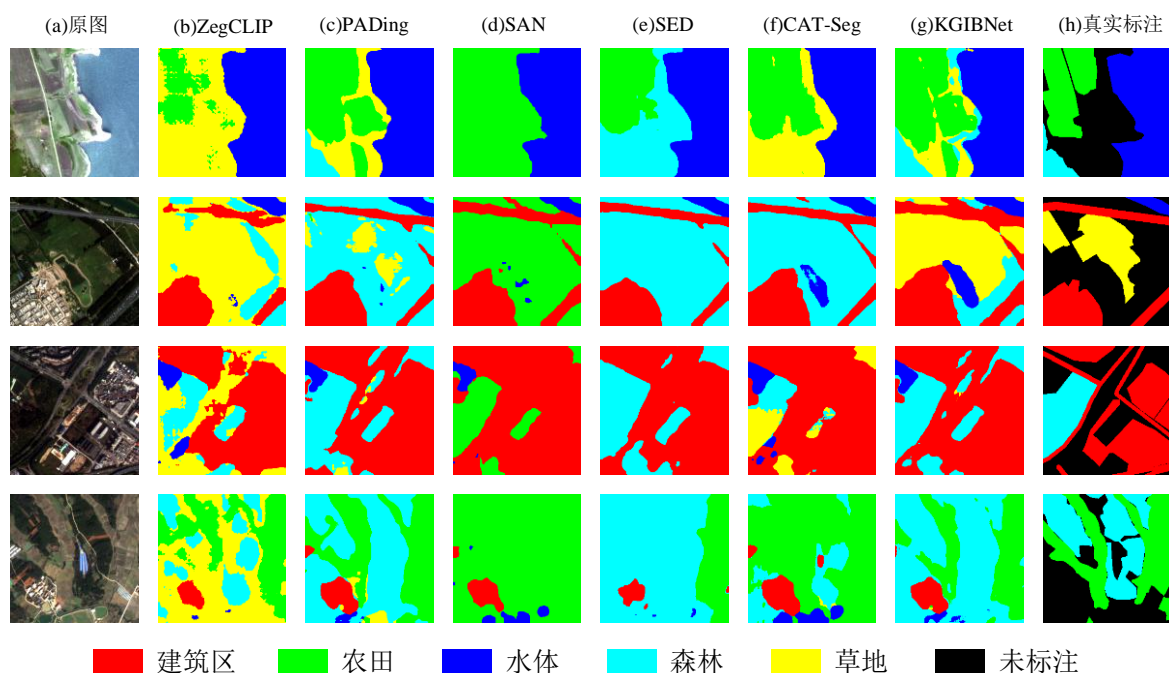


图 4.5 GID5 数据集上 KGIBNet 与不同方法的预测结果

Fig.4.5 Prediction results of KGIBNet and different methods on the GID5 dataset

如表 4.1 所示, KGIBNet 在 GID5 数据集上的整体性能优于其他方法, mIoU 和 mACC 分别达到 72.04%和 81.07%, 相较于 SAN 有显著提升。在不可见类别上, KGIBNet 的森林和草地 IoU 分别达到 84.73%和 29.81%, 远超 ZegCLIP、SED 以及 SAN。这一提升表明, 知识图谱的语义增强和跨模态融合策略显著提高了模型对未见类别的泛化能力。在可见

类别上，KGIBNet 同样表现出色，例如建筑区的 IoU 为 93.40%，水体的 IoU 为 76.23%，均优于其他方法。KGIBNet 通过知识图谱的引入增强了类别间的语义关系建模，提升了模型对不可见类别的泛化能力，同时保持了可见类别的分割精度，展现了其在零样本语义分割任务中的优越性。

表 4.1 GID5 数据集上 KGIBNet 与不同方法的精度比较

Table 4.1 Comparison of the accuracies between KGIBNet and different methods on the GID5 dataset

类别	ZegCLIP		PADing		SAN		SED		CAT-Seg		KGIBNet	
	I(%)	A(%)	I(%)	A(%)	I(%)	A(%)	I(%)	A(%)	I(%)	A(%)	I(%)	A(%)
建筑区	83.06	83.80	92.72	93.16	85.54	96.11	92.12	93.64	92.74	97.33	93.40	96.40
农田	57.20	59.96	71.23	80.12	45.46	98.39	78.13	84.20	81.35	91.21	76.01	81.07
水体	79.97	90.04	75.15	80.15	54.22	92.02	80.24	90.03	73.12	88.52	76.23	94.81
森林	77.99	80.12	81.56	91.12	34.14	34.48	78.25	96.77	75.20	83.22	84.73	96.54
草地	24.32	76.15	32.15	42.47	8.61	13.43	0.53	0.53	25.86	38.88	29.81	36.54
mIoU(%)	64.51		70.56		45.59		65.85		69.65		72.04	
mACC(%)	78.01		77.40		66.89		73.03		79.83		81.07	

注：表中 I 和 A 分别代表各类别的 IoU 和 ACC；双横线上的为可见类别；双横线下的为不可见类别；粗体代表最高精度值

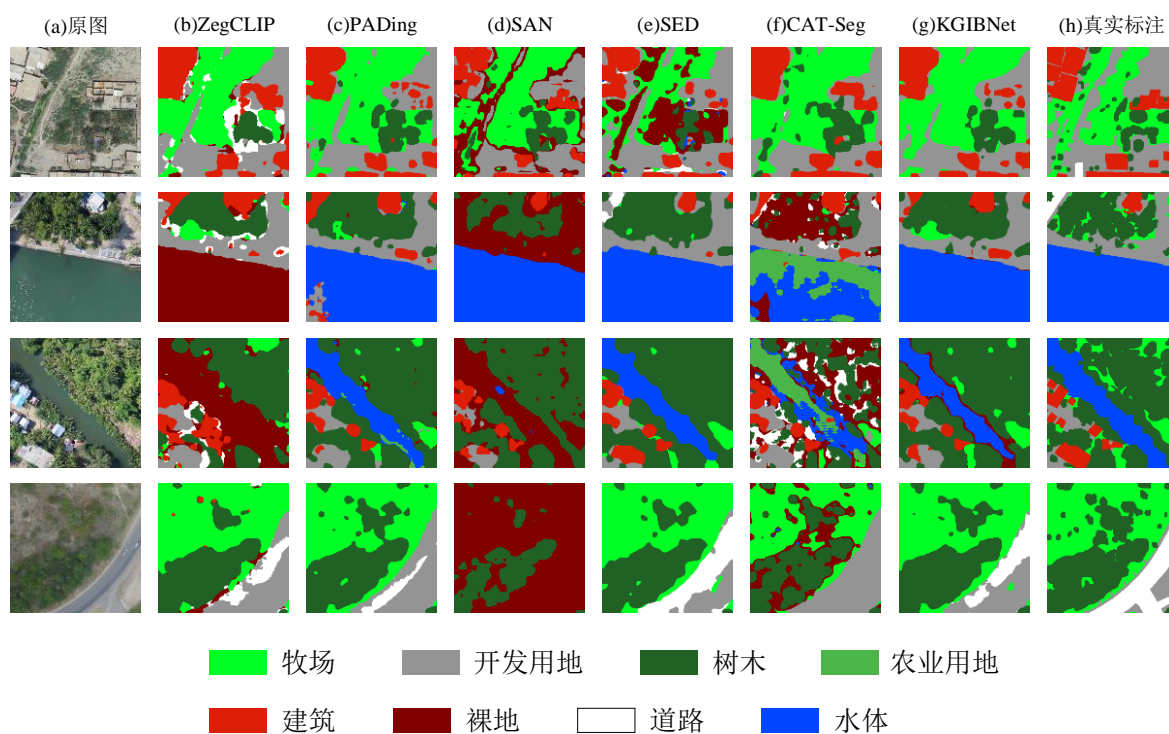


图 4.6 OpenEarthMap 数据集上 KGIBNet 与不同方法的预测结果

Fig.4.6 Prediction results of KGIBNet and different methods on the OpenEarthMap dataset

图 4.6 展示了不同方法在 OpenEarthMap 数据集上的预测结果，各模型在可见和不可见类别的分割性能上表现出显著差异。ZegCLIP 在处理不可见类别时表现不足，尤其对水体的识别准确率较低，并与裸地类别产生了严重的混淆。这一问题可能源于其基于文本-

图像块嵌入相似性的直接掩码生成策略，缺乏对语义相近类别的精细区分能力。PADing 方法在整体分割效果上表现较为优异，得益于其生成式特征合成策略，然而在水体区域的预测中出现了显著错误，将部分水体误判为开发用地。SAN 方法在可见和不可见类别的分割任务中均表现欠佳，类别混淆问题尤为突出，其分割效果不够理想。这可能与其冻结 CLIP 主干、仅依赖轻量侧网络的设计有关，限制了模型对复杂场景的适配能力。值得注意的是，SED 方法在不可见类别（如道路和水体）的分割任务中展现出较强的识别能力，其性能优于其他对比方法。相比之下，CAT-Seg 方法的表现稍逊于 SED，且存在将其他类别误分为裸地的倾向。在所有这些方法中，KGIBNet 通过引入知识图谱进行语义关系建模和错误纠正，不仅显著提升了不可见类别的分割精度，同时在可见类别上也保持了稳定的预测性能，最终实现了最优的整体分割效果，展现出更好的准确性和鲁棒性。

从表 4.2 数据来看，KGIBNet 在 OpenEarthMap 数据集上的整体性能优于其他方法，mIoU 和 mACC 分别达到 43.26%和 53.75%，超越了其他方法。在不可见类别（裸地、道路和水体）上，KGIBNet 的 IoU 分别为 2.40%、20.56%和 25.42%，显著优于 ZegCLIP、PADing 以及 SAN。SED 和 CAT-Seg 在不可见类别上的表现虽优于前三者，但仍不及 KGIBNet，例如 SED 在裸地的 IoU 仅为 0.05%，表明其对不可见类别的表征能力不足。在可见类别上，KGIBNet 同样表现出色，例如农业用地的 IoU 达到 65.71%，牧场的 IoU 为 51.14%，均优于其他方法。

表 4.2 OpenEarthMap 数据集上 KGIBNet 与不同方法的精度比较

Table 4.2 Comparison of the accuracies between KGIBNet and different methods on the OpenEarthMap dataset

类别	ZegCLIP		PADing		SAN		SED		CAT-Seg		KGIBNet	
	I(%)	A(%)	I(%)	A(%)	I(%)	A(%)	I(%)	A(%)	I(%)	A(%)	I(%)	A(%)
农业用地	58.96	61.51	54.26	83.21	60.92	77.91	58.10	78.68	60.00	77.34	65.71	76.46
牧场	43.16	51.17	49.91	63.07	19.26	21.22	48.32	66.87	50.55	68.45	51.14	64.87
开发用地	40.53	65.57	42.29	74.51	20.42	24.57	41.96	72.12	42.14	72.81	51.17	62.37
建筑	66.08	82.94	56.15	71.15	40.72	44.11	69.34	82.18	71.27	84.89	68.46	80.56
树木	56.12	63.58	65.80	83.15	62.89	71.76	64.95	79.49	64.39	81.80	61.23	77.55
裸地	4.17	34.73	0.05	0.09	2.83	72.32	0.05	0.06	2.97	4.48	2.40	9.12
道路	10.13	12.88	2.08	2.65	3.36	3.42	15.45	17.25	7.05	7.23	20.56	23.87
水体	0.09	0.15	9.54	9.59	21.62	22.41	30.83	31.32	23.93	24.02	25.42	35.15
mIoU(%)	34.91		35.01		29.00		41.12		40.29		43.26	
mACC(%)	46.57		48.43		42.22		53.50		52.63		53.75	

注：表中 I 和 A 分别代表各类别的 IoU 和 ACC；双横线上的为可见类别；双横线下的为不可见类别；粗体代表最高精度值

图 4.7 展示了不同方法在 LoveDA 数据集上的预测结果。ZegCLIP 和 PADing 在整体分割效果上表现较为优异，得益于其充分利用了 CLIP 的预训练特征。然而，二者在裸地和农业用地等类别间出现了明显的混淆现象。SAN 的预测效果尤为欠佳，甚至出现整张图像被错误预测为单一类别的情况，显示出其在类别区分能力上的严重不足。这一现象可能与其仅依赖轻量侧网络的设计有关，限制了模型对遥感场景的适配能力。相比之下，SED 和 CAT-Seg 相较于两阶段方法有了显著改进，但 SED 在不可见类别裸地的识别上效

果较差，而 CAT-Seg 则难以精确提取各类别的边界，分割结果仍存在模糊性。相比之下，KGIBNet 通过知识图谱的语义纠正，显著提升了不可见类别（如水体）的提取精度，同时在农业用地等可见类别的分割上也表现出色，展现了更强的边界提取能力和整体分割性能。

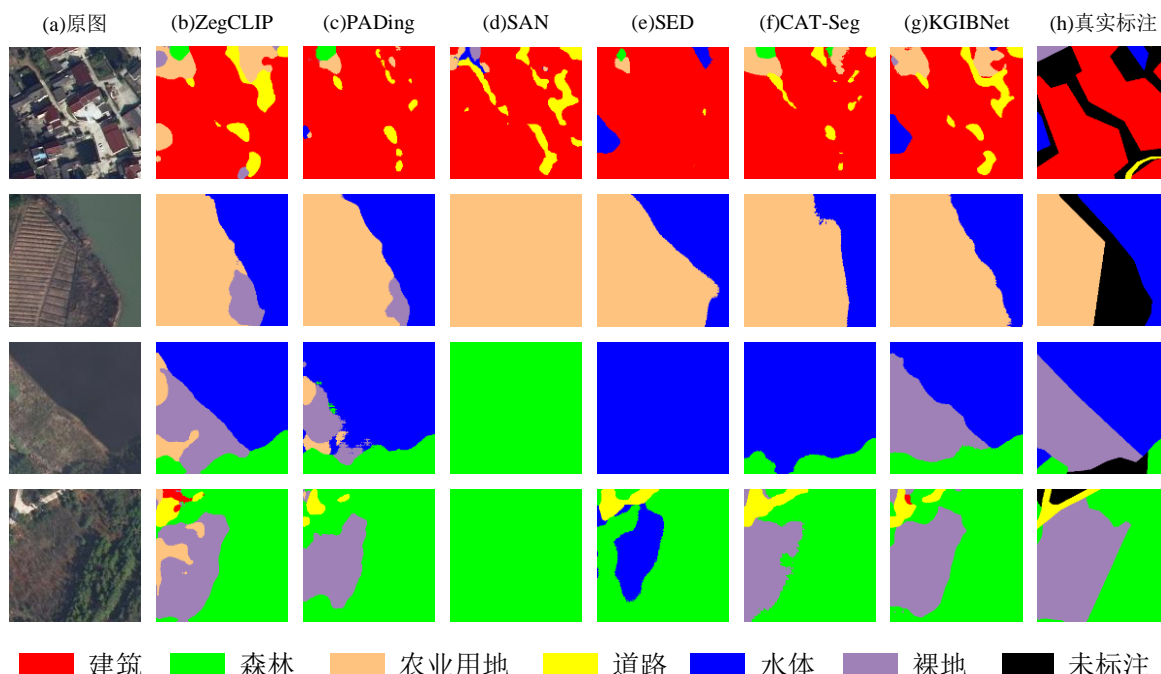


图 4.7 LoveDA 数据集上 KGIBNet 与不同方法的预测结果

Fig.4.7 Prediction results of KGIBNet and different methods on the LoveDA dataset

表 4.3 LoveDA 数据集上 KGIBNet 与不同方法的精度比较

Table 4.3 Comparison of the accuracies between KGIBNet and different methods on the LoveDA dataset

类别	ZegCLIP		PADing		SAN		SED		CAT-Seg		KGIBNet	
	I(%)	A(%)	I(%)	A(%)	I(%)	A(%)	I(%)	A(%)	I(%)	A(%)	I(%)	A(%)
建筑	68.54	92.65	51.98	56.15	65.16	88.99	74.75	90.55	70.57	92.34	73.94	88.66
道路	65.14	81.26	71.15	81.65	66.22	74.37	74.70	84.05	68.06	75.12	73.19	85.97
森林	62.45	71.56	58.16	68.17	33.11	70.28	59.80	67.67	63.75	78.36	63.74	75.91
农业用地	52.12	95.77	46.05	53.16	53.47	86.80	61.71	93.36	53.58	95.23	55.44	96.34
水体	13.16	15.15	15.56	21.16	15.58	16.15	38.70	57.80	11.99	12.01	17.77	18.07
裸地	20.19	23.18	15.47	31.25	15.68	16.74	1.50	1.51	35.80	48.16	34.88	48.49
mIoU(%)	46.93		43.06		41.54		51.86		50.63		53.16	
mACC(%)	63.26		51.92		58.89		65.82		66.87		68.91	

注：表中 I 和 A 分别代表各类别的 IoU 和 ACC；双横线上的为可见类别；双横线下的为不可见类别；粗体代表最高精度值

根据表 4.3，KGIBNet 在 LoveDA 数据集上的整体性能表现最佳，mIoU 和 mACC 分别达到 53.16%和 68.91%。在不可见类别的分割上，KGIBNet 的 IoU 分别为 17.77%和 34.88%，相较于 ZegCLIP（46.93%和 63.26%）、PADing（43.06%和 51.92%）和 SAN（15.58%和 15.68%）展现了更强的泛化能力。SED 和 CAT-Seg 在不可见类别上的表现有所提升，但 SED 在裸地的 IoU 仅为 1.50%，而 CAT-Seg 在水体的 IoU 仅为 11.99%，均不

及 KGIBNet。在可见类别上，KGIBNet 同样表现出色，例如建筑的 IoU 为 73.94%，道路的 IoU 为 73.19%，森林的 IoU 为 63.74%，均优于其他方法。

4.3.2 消融实验分析

本节通过消融实验验证了 KGIBNet 各模块在零样本语义分割任务中的作用，实验以仅使用 CLIP 视觉-文本嵌入计算语义匹配特征图并输入分割头的方案作为基线，旨在评估各模块的独立贡献及协同效应。实验在 GID5、OpenEarthMap 和 LoveDA 三个数据集上进行，结果如表所示。基线方法的 mIoU 分别为 61.49%、35.15%和 42.46%，mACC 分别为 76.56%、48.16%和 58.15%。

引入属性知识增强模块（AKEM）后，通过融合知识图谱中的地物类别属性信息，模型的语义区分能力显著提升。例如，在 GID5 数据集上 mIoU 增至 64.56%，在 OpenEarthMap 上 mACC 从 48.16%提升至 54.15%，显示出知识嵌入对语义理解的增强作用。进一步加入基于信息瓶颈的知识融合模块（KFBIB）后，通过筛选与模态融合相关的信息并滤除噪声，模型性能得到提升。例如，在 GID5 数据集上 mIoU 升至 70.51%，在 LoveDA 数据集上 mACC 达到 68.84%，表明 KFBIB 有效优化了跨模态特征的融合质量，提升了分割精度和鲁棒性。

当 AKEM 和 KFBIB 模块协同作用时，KGIBNet 在三个数据集上的性能达到最佳，mIoU 分别提升至 72.04%、42.31%和 53.16%，mACC 为 81.07%、53.77%和 68.91%。与基线相比，GID5 上的 mIoU 提升了 10.55 个百分点，LoveDA 上的 mACC 提高了 10.76 个百分点。这一结果充分证明，AKEM 通过知识图谱增强了语义表达，KFBIB 通过信息瓶颈优化了特征融合，两者协同作用显著提升了 KGIBNet 对不可见类别的区分能力和整体分割性能，为遥感图像零样本语义分割提供了创新的解决方案。

表 4.4 KGIBNet 的消融实验

Table 4.4 The abalition experiment of KGIBNet

方法	GID5		OpenEarthMap		LoveDA	
	mIoU(%)	mACC(%)	mIoU(%)	mACC(%)	mIoU(%)	mACC(%)
基线方法	61.49	76.56	35.15	48.16	42.46	58.15
基线方法+AKEM	64.56	78.17	38.18	54.15	48.16	62.16
基线方法+KFBIB	70.51	83.12	40.65	50.14	52.23	68.84
基线方法+AKEM+KFBIB	72.04	81.07	42.31	53.77	53.16	68.91

注：粗体代表最高精度值

4.3.3 知识增强嵌入可视化分析

为了验证属性知识增强模块在提升类别语义区分能力方面的有效性，本研究采用 t-SNE 降维可视化方法，对原始类别文本嵌入与知识增强嵌入展开对比分析。图 4.8 呈现了不同嵌入的可视化结果。

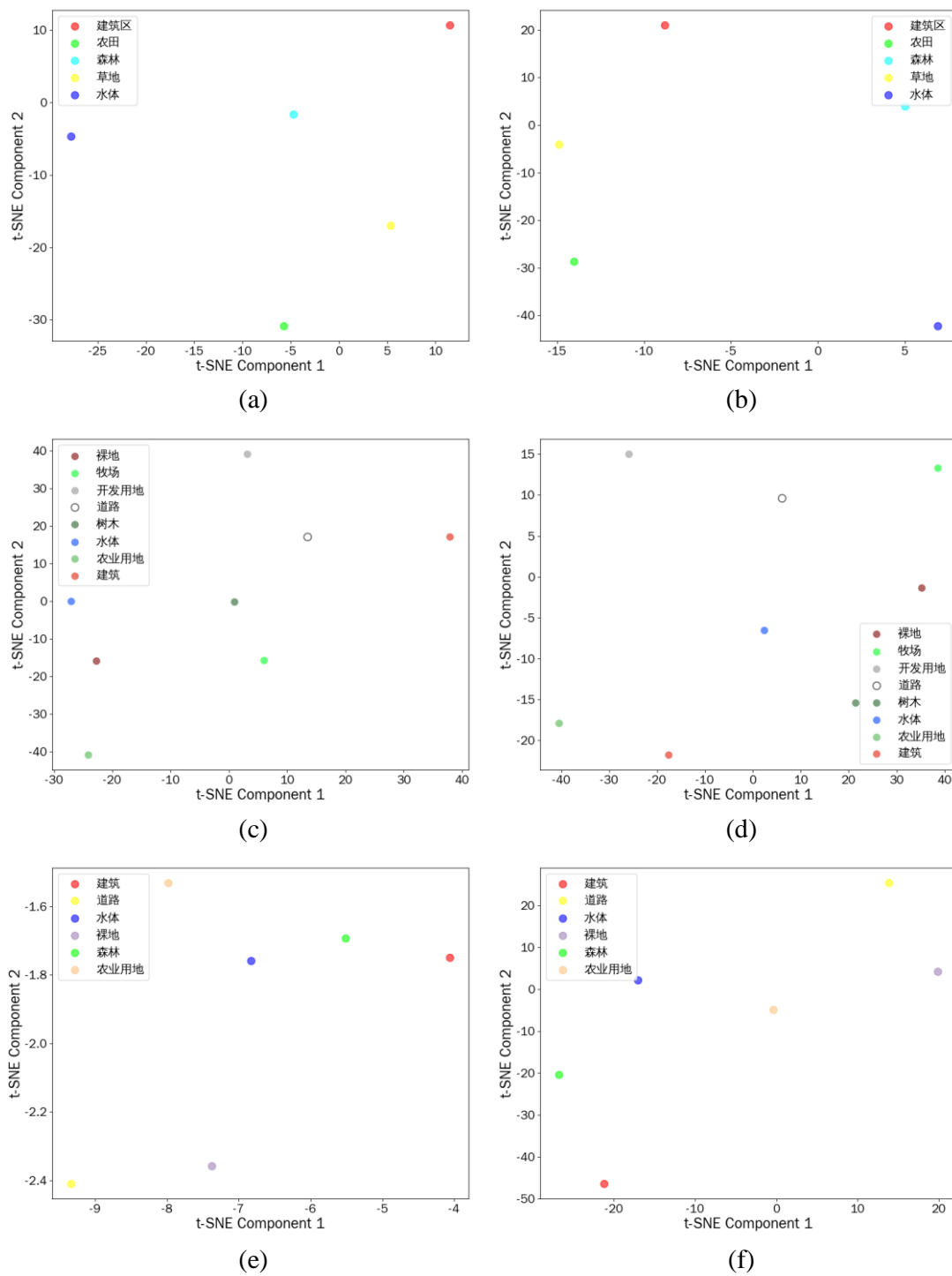


图 4.8 原始类别文本嵌入和知识增强嵌入的可视化比较

Fig.4.8 Visual Comparison of the Original Categorical Text Embeddings and the Knowledge-Enhanced Embeddings

其中子图 (a) 和 (b) 分别展示了在 GID 数据集上知识增强嵌入和原始类别文本嵌入的 t-SNE 可视化情况。从数据分布的直观角度来看, 在 GID 数据集中, 原始类别文本嵌入下“森林”与“农田”的空间分布相对较为离散, 表明二者语义距离较远, 语义关联性模糊。而引入属性知识增强模块后, 融入地学知识使得“森林”与“农田”在可视化空间

中的位置显著靠近。这背后的原理在于，地学知识为模型提供了关于土地覆盖类型间生态、地理关联的先验信息，从而引导模型在学习嵌入时拉近二者语义距离，增强语义关联性。

子图 (c) 和 (d) 展示了在 OpenEarthMap 数据集上知识增强嵌入和原始类别文本嵌入的可视化结果。在 OpenEarthMap 数据集中，不同领域来源的类别名称存在明显的领域特异性，这给模型统一理解和处理类别语义带来挑战。原始类别文本嵌入因缺乏对这种领域特异性的有效应对机制，导致语义表达较为局限。属性知识增强模块通过语义建模，将知识图谱中的三元组信息（如“建筑 - 临近 - 道路”）融入原始类别嵌入。这一过程实质是为模型提供了丰富的语义约束与关联线索，让模型能够从更通用的视角理解类别间关系，有效缓解了类别名称的领域特异性问题。

子图 (e) 和 (f) 分别展示了在 LoveDA 数据集上知识增强嵌入和原始类别文本嵌入的可视化结果。在 LoveDA 数据集下，子图 (f) 中原始嵌入呈现出较为均匀的分布状态，这意味着各类别在语义空间中缺乏明显的区分特征，难以准确识别类别间差异。反观知识增强嵌入，成功捕捉到“森林”和“农业用地”靠近“水体”的空间语义关系。这是因为属性知识增强模块在处理过程中，能够利用知识图谱中关于地理要素的属性知识，对原始类别嵌入进行优化，使得模型输出的嵌入结果能够更好地反映现实世界中地理类别间的依存关系，提升了类别的语义区分能力。

属性知识增强模块通过融合知识图谱的结构化信息，不仅优化了类别语义表示，还增强了模型在不同遥感数据集上的语义区分能力，为零样本语义分割任务奠定了更具鲁棒性的语义嵌入基础。

4.3.4 超参数分析

从 GID 数据集的实验结果来看，当 $\lambda = 0.5, \beta = 0.5$ 时，mIoU 达到最高值 73.00%，表明该超参数组合使模型在 GID 数据集上表现出最佳性能。然而， β 在 0.5 和 1.0 之间的变化对结果影响较小，而当 β 增至 2.0 时，mIoU 显著下降，例如在 $\lambda = 0.5$ 时降至 65.23%。这可能是由于过高的 β 值导致模型对冗余信息的抑制过于强烈，从而丢失了部分关键的语义信息。在 OpenEarthMap 数据集上，当 $\lambda = 0.5, \beta = 1.0$ 时，mIoU 达到最高值 43.26%，但整体性能波动较大，例如 $\beta = 2.0$ 时 mIoU 下降至 36.43%，显示出该数据集对超参数变化的敏感性较高。在 LoveDA 数据集上的结果表明， $\lambda = 0.5, \beta = 1.0$ 时 mIoU 达到最高值 53.16%，而当 $\beta = 0.5$ 时性能明显下降，例如在 $\lambda = 1.0$ 时降至 48.92%。这可能反映出较低的 β 值未能充分抑制冗余信息，从而干扰了模型的语义区分能力。综合上述分析，选择 $\beta = 1.0$ 作为较优参数。

λ 的取值同样对模型性能产生显著影响。在固定 $\beta = 1.0$ 的情况下，当 $\lambda = 0.1$ 时，信息瓶颈优化损失的权重过低，导致模型过于依赖交叉熵损失，未能充分利用知识图谱的空间和属性信息优化。例如，在 OpenEarthMap 数据集上，mIoU 仅为 41.49%。随着 λ

增加至 0.5，模型在 GID、OpenEarthMap 和 LoveDA 数据集上的 mIoU 分别提升至 72.04%、43.26% 和 53.16%，表明信息瓶颈优化显著增强了语义分割性能。然而，当 λ 进一步增至 1.0 时，mIoU 出现下降趋势，例如在 GID 数据集上降至 68.48%。这可能是因为过高的 λ 值使模型过于聚焦信息瓶颈优化，而牺牲了像素级分类的精确性。

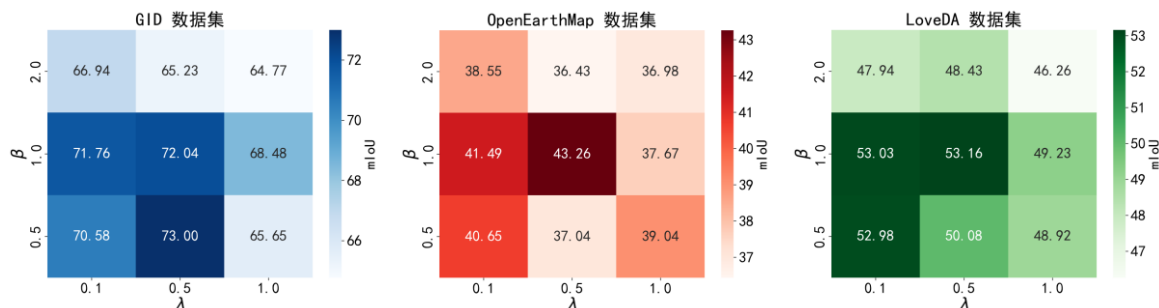


图 4.9 不同数据集的超参数分析

Fig.4.9 Hyperparameter Analysis of Different Datasets

综合来看， $\lambda = 0.5$ 和 $\beta = 1.0$ 的超参数组合能够有效平衡语义优化与分类精度，为零样本语义分割任务提供了最佳支持。

4.3.5 模型复杂度分析

本实验对 ZegCLIP、SAN、SED、Cat-Seg 和 KGIBNet 在 GID5 数据集上进行了比较，聚焦关键复杂度指标并结合性能进行分析。

表 4.5 模型复杂度与性能对比

Table 4.5 Model Complexity and Performance Comparison

相关指标	zegclip	SAN	SED	Cat-Seg	KGIBNet
可训练参数 (M)	13.8	33.7	89.6	70.3	30.6
训练速度 (s/iter)	0.2069	0.312	0.3672	0.5036	0.3406
推理速度 (s/iter)	0.0209	0.1038	0.0054	0.0217	0.0052
mIoU (%)	64.51	45.59	65.85	69.65	72.04

注：粗体代表最优值

ZegCLIP 以最少的参数量和最快的训练速度展现出轻量化优势，适合资源受限场景，但推理效率和性能有所下降；SAN 模型参数较少，但推理速度较慢，实用性受限；SED 在推理速度上表现优异，但参数量最大，增加了计算负担；Cat-Seg 训练速度最慢，复杂度偏高，适合精度优先的应用；我们的 KGIBNet 以适中的参数量（30.6M）和最快的推理速度（0.0052 s/iter）脱颖而出。从实验结果来看，KGIBNet 在复杂度控制上表现出色，其推理效率超越所有对比模型。相较于 SED 的高参数负担和 Cat-Seg 的训练开销，KGIBNet 在资源需求与运行速度间达到优异平衡，展现出更高的部署潜力和应用价值。

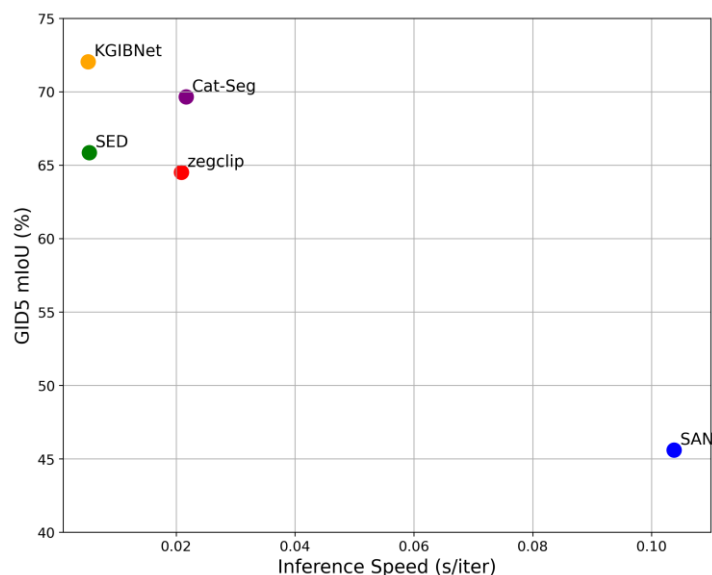


图 4.10 GID5 数据集不同方法的 mIoU 和推理速度对比

Fig.4.10 Comparison of mIoU and Inference Speed of Different Methods on the GID5 Dataset

4.4 本章小结 (Summary)

本章提出了基于知识图谱和信息瓶颈理论的零样本语义分割方法 KGIBNet，并在 GID5、OpenEarthMap 和 LoveDA 三个遥感数据集上进行了全面验证。KGIBNet 通过属性知识增强模块 (AKEM) 和基于信息瓶颈的知识融合模块 (KFBIB)，有效融合知识图谱的结构化信息，优化了类别语义嵌入和视觉特征表示，显著提升了不可见类别的分割性能。实验结果表明，KGIBNet 在 GID5 数据集上的 mIoU 和 mACC 分别达到 72.04% 和 81.07%，在 OpenEarthMap 和 LoveDA 数据集上也取得了 43.26%/53.75% 和 53.16%/68.91% 的优异成绩，超越了 ZegCLIP、PADing、SAN、SED 和 CAT-Seg 等方法。消融实验进一步验证了 AKEM 和 KFBIB 模块的协同作用，证明其在增强语义区分能力和减少冗余信息方面的关键贡献。KGIBNet 为遥感图像零样本语义分割提供了高效、鲁棒的解决方案，展现了其在复杂地物分割任务中的应用潜力。

5 总结与展望

5 Conclusion and Prospect

5.1 总结 (Conclusion)

本文针对遥感图像零样本语义分割中的关键挑战——多尺度目标分布、旋转不变性不足以及地物类别的领域特异性语义问题，提出了一系列创新方法，并在 GID5、OpenEarthMap 和 LoveDA 三个数据集上进行了全面验证。研究围绕两个核心模型展开：多尺度增强和旋转不变性网络 (MSRINet) 以及基于知识图谱和信息瓶颈理论的零样本语义分割网络 (KGIBNet)，分别从视觉特征优化和语义知识增强两个维度提供了解决方案。

MSRINet 通过全局-局部信息增强模块 (GLIEM)、旋转不变性特征优化模块 (RIFOM) 和类别关系建模模块 (CRMM)，有效应对了遥感图像中目标尺度差异大和多视角成像带来的挑战。GLIEM 融合全局与局部特征构建多尺度表示，RIFOM 通过旋转等变特征提取提升模型鲁棒性，而 CRMM 则通过自注意力机制强化类别间语义关联。实验结果表明，MSRINet 在 GID5 数据集上的 mIoU 和 mACC 分别达到 72.45% 和 84.40%，在不可见类别（如森林和草地）的分割任务中表现出色，IoU 分别达到 88.94% 和 35.28%，显著优于 ZegFormer、SAN、SED 和 CAT-Seg 等方法。消融实验进一步验证了各模块的协同作用，凸显了全部-局部特征融合和旋转不变性优化对提升分割性能的关键贡献。

在此基础上，KGIBNet 进一步针对地物类别的领域特异性语义和视觉特征复杂性问题，通过属性知识增强模块 (AKEM) 和基于信息瓶颈的知识融合模块 (KFBIB) 实现了性能突破。AKEM 利用知识图谱的属性信息生成增强的语义嵌入，KFBIB 则通过图卷积网络和信息瓶颈理论优化多模态特征融合，滤除冗余信息。实验结果显示，KGIBNet 在 GID5 数据集上的 mIoU 和 mACC 分别达到 72.04% 和 81.07%，在 OpenEarthMap 和 LoveDA 数据集上分别取得 43.26%/53.75% 和 53.16%/68.91% 的优异成绩，尤其在不可见类别的分割中表现出更强的适应性。消融实验证实了 AKEM 和 KFBIB 模块在提升语义区分能力和减少模态间冗余信息的核心作用。

MSRINet 与 KGIBNet 之间存在紧密联系与互补性：MSRINet 通过视觉特征的优化奠定了零样本分割的基础，解决了多尺度与多视角问题；而 KGIBNet 在此之上引入知识图谱和信息瓶颈理论，进一步增强了语义理解与特征精炼能力，弥补了 MSRINet 在领域特异性语义表达上的不足。两者结合，形成了从特征提取到语义增强的完整技术链条，成功应对了遥感图像零样本语义分割中的多重挑战，为复杂地物分割任务提供了高效、鲁棒的解决方案，展现了其在城市规划、环境监测等领域的广阔应用前景。

5.2 展望 (Prospect)

尽管本文提出的 MSRINet 和 KGIBNet 在遥感图像零样本语义分割任务中取得了显著成果,但仍存在一些不足之处,未来研究可从以下几个方面进一步深化。首先,MSRINet 和 KGIBNet 在多尺度特征融合和知识图谱推理过程中引入了额外的计算开销,尤其在处理高分辨率遥感图像时,计算复杂度和内存需求较高,可探索轻量化模型设计或高效计算策略(如模型剪枝、量化或知识蒸馏),以提升模型在资源受限环境下的适用性。其次,本文使用的知识图谱主要基于 RSKG 和 ConceptNet,知识范围有限且构建更新依赖人工标注,难以适应快速变化的遥感场景,未来可研究自动化知识图谱构建方法,如通过大语言模型从遥感文献中提取知识,或设计动态更新机制以适应新地物类别和场景。此外,本文主要利用 CLIP 的视觉-文本对齐能力和知识图谱的语义信息,未来可进一步引入光谱信息、时序信息或地理元数据等其他模态数据,通过多模态融合技术提升模型对复杂地物类别的理解能力,尤其是在光谱特征相似但语义不同的类别区分上。同时,尽管模型在 GID5、OpenEarthMap 和 LoveDA 数据集上表现优异,但其在更广泛的遥感场景(如极地、沙漠或高山地区)中的泛化能力仍需验证,未来可扩展实验范围,测试模型在多样化场景下的鲁棒性,并探索其在实际应用(如灾害实时监测、生态保护)中的部署可行性。最后,当前模型的决策过程较为复杂,缺乏对分割结果的可解释性分析,可引入可解释性方法(如注意力可视化或语义推理路径分析)增强模型可信度,同时开发用户交互界面,允许用户根据实际需求调整模型预测结果,以提升其实用性。通过以上方向的深入研究,期望进一步提升遥感图像零样本语义分割技术的性能和应用价值,为遥感领域的智能化分析和决策支持提供更强大的技术支撑。

参考文献

- [1] Yuan X, Shi J, Gu L. A review of deep learning methods for semantic segmentation of remote sensing imagery[J]. Expert Systems with Applications, 2021, 169: 114417.
- [2] Huang L, Jiang B, Lv S, et al. Deep-learning-based semantic segmentation of remote sensing images: A survey[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2023, 17: 8370-8396.
- [3] 李德仁, 李明. 无人机遥感系统的研究进展与应用前景 [J]. 武汉大学学报 (信息科学版), 2014, 39(05): 505-513+540. DOI:10.13203/j.whugis.20140045.
- [4] 刘纪远, dls.iogac.on, 布和敖斯尔. 中国土地利用变化现代过程时空特征的研究——基于卫星遥感数据[J]. 第四纪研究, 2000, (03): 229-239.
- [5] Cheng J, Deng C, Su Y, et al. Methods and datasets on semantic segmentation for Unmanned Aerial Vehicle remote sensing images: A review[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2024, 211: 1-34.
- [6] Pourpanah F, Abdar M, Luo Y, et al. A review of generalized zero-shot learning methods[J]. IEEE transactions on pattern analysis and machine intelligence, 2022, 45(4): 4051-4070.
- [7] Romera-Paredes B, Torr P. An embarrassingly simple approach to zero-shot learning[C]//International conference on machine learning. PMLR, 2015: 2152-2161.
- [8] Wang W, Zheng V W, Yu H, et al. A survey of zero-shot learning: Settings, methods, and applications[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2019, 10(2): 1-37.
- [9] Li Y, Zhu Z, Yu J G, et al. Learning deep cross-modal embedding networks for zero-shot remote sensing image scene classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 59(12): 10590-10603.
- [10] Li Y, Kong D, Zhang Y, et al. Robust deep alignment network with remote sensing knowledge graph for zero-shot and generalized zero-shot remote sensing image scene classification[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2021, 179: 145-158.
- [11] Bucher M, Vu T H, Cord M, et al. Zero-shot semantic segmentation[J]. Advances in Neural Information Processing Systems, 2019, 32.
- [12] Ding J, Xue N, Xia G S, et al. Decoupling zero-shot semantic segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 11583-11592.
- [13] Zhou Z, Lei Y, Zhang B, et al. Zegclip: Towards adapting clip for zero-shot semantic segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 11175-11185.
- [14] Diakogiannis F I, Waldner F, Caccetta P, et al. ResUNet-a: A deep learning framework for semantic

- segmentation of remotely sensed data[J]. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2020, 162: 94-114.
- [15] 崔宾阁, 吴景, 李心慧, 任广波, 路燕.2023.结合深度学习和植被指数的滨海湿地高分二号遥感影像信息提取.遥感学报, 27 (6): 1376-1386 DOI: 10.11834/jrs.20221658.
- [16] Li R, Zheng S, Zhang C, et al. Multiattention network for semantic segmentation of fine-resolution remote sensing images[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2021, 60: 1-13.
- [17] Zhao Q, Liu J, Li Y, et al. Semantic segmentation with attention mechanism for remote sensing images[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2021, 60: 1-13.
- [18] Ma X, Zhang X, Pun M O, et al. A Multilevel Multimodal Fusion Transformer for Remote Sensing Semantic Segmentation[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2024, 62: 3373033.
- [19] Li X, Wen C, Hu Y, et al. RS-CLIP: Zero shot remote sensing scene classification via contrastive vision-language supervision[J]. *International Journal of Applied Earth Observation and Geoinformation*, 2023, 124: 103497.
- [20] Sumbul G, Cinbis R G, Aksoy S. Fine-grained object recognition and zero-shot learning in remote sensing imagery[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2017, 56(2): 770-779.
- [21] Li A, Lu Z, Wang L, et al. Zero-shot scene classification for high spatial resolution remote sensing images[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2017, 55(7): 4157-4167.
- [22] Chaudhuri U, Bose R, Banerjee B, et al. Zero-shot cross-modal retrieval for remote sensing images with minimal supervision[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, 60: 1-15.
- [23] Huang S, He S, Wen B. ZoRI: Towards Discriminative Zero-Shot Remote Sensing Instance Segmentation[J]. *arXiv preprint arXiv:2412.12798*, 2024.
- [24] Bane D, Singh A, Suri M. RS-A2M: Zero-shot Prompt with Attention-based Unsupervised Segmentation for Remote Sensing[C]//2024 IEEE Space, Aerospace and Defence Conference (SPACE). IEEE, 2024: 120-123.
- [25] Li K, Liu R, Cao X, et al. Segearth-ov: Towards training-free open-vocabulary segmentation for remote sensing images[J]. *arXiv preprint arXiv:2410.01768*, 2024.
- [26] Cao Q, Chen Y, Ma C, et al. Open-vocabulary remote sensing image semantic segmentation[J]. *arXiv preprint arXiv:2409.07683*, 2024.
- [27] Huang B, Li J, Luan W, et al. Expanding Open-Vocabulary Understanding for UAV Aerial Imagery: A Vision–Language Framework to Semantic Segmentation[J]. *Drones*, 2025, 9(2): 155.
- [28] Li W, Chen K, Chen H, et al. Geographical knowledge-driven representation learning for remote sensing images[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2021, 60: 1-16.
- [29] Li Y, Zhang Y. A new paradigm of remote sensing image interpretation by coupling knowledge graph and deep learning[J]. *Geomatics and Information Science of Wuhan University*, 2022, 47(8): 1176-1190.

- [30] Gun Z, Chen J. Novel knowledge graph-and knowledge reasoning-based classification prototype for OBIA using high resolution remote sensing imagery[J]. *Remote Sensing*, 2023, 15(2): 321.
- [31] Cui W, Yao M, Hao Y, et al. Knowledge and geo-object based graph convolutional network for remote sensing semantic segmentation[J]. *Sensors*, 2021, 21(11): 3848.
- [32] Wang F, Zhu X, Cheng X, et al. Mmkgat: Multi-modal knowledge graph-aware deep graph attention network for remote sensing image recommendation[J]. *Expert Systems with Applications*, 2024, 235: 121278.
- [33] Li Z, Liu F, Yang W, et al. A survey of convolutional neural networks: analysis, applications, and prospects[J]. *IEEE transactions on neural networks and learning systems*, 2021, 33(12): 6999-7019.
- [34] O'shea K, Nash R. An introduction to convolutional neural networks[J]. *arXiv preprint arXiv:1511.08458*, 2015.
- [35] Ajit A, Acharya K, Samanta A. A review of convolutional neural networks[C]//2020 international conference on emerging trends in information technology and engineering (ic-ETITE). IEEE, 2020: 1-5.
- [36] Chen Y, Dai X, Liu M, et al. Dynamic relu[C]//European conference on computer vision. Cham: Springer International Publishing, 2020: 351-367.
- [37] Murray N, Perronnin F. Generalized max pooling[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 2473-2480.
- [38] Bieder F, Sandkühler R, Cattin P C. Comparison of methods generalizing max-and average-pooling[J]. *arXiv preprint arXiv:2103.01746*, 2021.
- [39] Pinkus A. Approximation theory of the MLP model in neural networks[J]. *Acta numerica*, 1999, 8: 143-195.
- [40] Wu Z, Shen C, Van Den Hengel A. Wider or deeper: Revisiting the resnet model for visual recognition[J]. *Pattern recognition*, 2019, 90: 119-133.
- [41] Han K, Xiao A, Wu E, et al. Transformer in transformer[J]. *Advances in neural information processing systems*, 2021, 34: 15908-15919.
- [42] Kirillov A, Girshick R, He K, et al. Panoptic feature pyramid networks[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 6399-6408.
- [43] Han K, Wang Y, Chen H, et al. A survey on vision transformer[J]. *IEEE transactions on pattern analysis and machine intelligence*, 2022, 45(1): 87-110.
- [44] Gibbons F X. Self-attention and behavior: A review and theoretical update[J]. *Advances in experimental social psychology*, 1990, 23: 249-303.
- [45] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. PmLR, 2021: 8748-8763.
- [46] Qian Q, Hu J. Online zero-shot classification with clip[C]//European Conference on Computer Vision.

- Cham: Springer Nature Switzerland, 2024: 462-477.
- [47] Hao X, Ji Z, Li X, et al. Construction and application of a knowledge graph[J]. Remote Sensing, 2021, 13(13): 2511.
- [48] Wang F, Zhu X, Cheng X, et al. Mmkgat: Multi-modal knowledge graph-aware deep graph attention network for remote sensing image recommendation[J]. Expert Systems with Applications, 2024, 235: 121278.
- [49] Speer R, Chin J, Havasi C. Conceptnet 5.5: An open multilingual graph of general knowledge[C]//Proceedings of the AAAI conference on artificial intelligence. 2017, 31(1).
- [50] Zhang S, Tong H, Xu J, et al. Graph convolutional networks: a comprehensive review[J]. Computational Social Networks, 2019, 6(1): 1-23.
- [51] Saxe A M, Bansal Y, Dapello J, et al. On the information bottleneck theory of deep learning[J]. Journal of Statistical Mechanics: Theory and Experiment, 2019, 2019(12): 124020.
- [52] Slonim N. The information bottleneck: Theory and applications[D]. Hebrew University of Jerusalem, 2002.
- [53] Kawaguchi K, Deng Z, Ji X, et al. How does information bottleneck help deep learning?[C]//International Conference on Machine Learning. PMLR, 2023: 16049-16096.
- [54] Blumenstiel B, Jakubik J, Kühne H, et al. What a mess: Multi-domain evaluation of zero-shot semantic segmentation[J]. Advances in Neural Information Processing Systems, 2023, 36: 73299-73311.
- [55] Tong X Y, Xia G S, Lu Q, et al. Land-cover classification with high-resolution remote sensing images using transferable deep models[J]. Remote Sensing of Environment, 2020, 237: 111322.
- [56] Xia J, Yokoya N, Adriano B, et al. Openearthmap: A benchmark dataset for global high-resolution land cover mapping[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023: 6254-6264.
- [57] Wang J, Zheng Z, Ma A, et al. LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation[J]. arXiv preprint arXiv:2110.08733, 2021.
- [58] Xu M, Zhang Z, Wei F, et al. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 736-753.
- [59] Xu M, Zhang Z, Wei F, et al. Side adapter network for open-vocabulary semantic segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 2945-2954.
- [60] Xie B, Cao J, Xie J, et al. Sed: A simple encoder-decoder for open-vocabulary semantic segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 3426-3436.
- [61] Cho S, Shin H, Hong S, et al. Cat-seg: Cost aggregation for open-vocabulary semantic

segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
2024: 4113-4123.

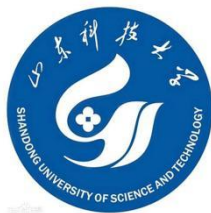
- [62] He S, Ding H, Jiang W. Primitive generation and semantic-related alignment for universal zero-shot segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.
2023: 11238-11247.

作者简历

致 谢

学位论文数据集

关键词*	密级*	中图分类号*	UDC	论文资助
遥感图像；零样本语义分割；多尺度特征；旋转不变性；知识图谱；信息瓶颈理论	公开	TP751	004	
学位授予单位名称*	学位授予单位代码*	学位类别*	学位级别*	
山东科技大学	10424	工学	工学硕士	
论文题名*		并列题名*		论文语种*
基于多尺度旋转不变与知识图谱优化的零样本遥感图像语义分割		Zero-shot Semantic Segmentation of Remote Sensing Images Based on Multi-scale Rotation Invariance and Knowledge Graph Optimization		中文
作者姓名*		学号*		
培养单位名称*	培养单位代码*	培养单位地址	邮编	
山东科技大学	10424	山东省青岛市	266590	
学科专业*	研究方向*	学制*	学位授予年*	
计算机应用技术	人工智能	三年	2025	
论文提交日期*		2025 年 月 日		
导师姓名*		职称*		
评阅人	答辩委员会主席*	答辩委员会成员		
电子版论文提交格式 文本（√） 图像（ ） 视频（ ） 音频（ ） 多媒体（ ） 其他（ ）				
推荐格式：application/msword; application/pdf				
电子版论文出版（发布）者	电子版论文出版（发布）地		权限声明	
论文总页数*				
注：共 33 项，其中带*为必填数据，共 22 项。				



工学硕士学位论文

基于多尺度旋转不变与知识图谱优化的零样本遥感图像 语义分割

Zero-shot Semantic Segmentation of Remote Sensing Images
Based on Multi-scale Rotation Invariance and Knowledge Graph
Optimization

作 者

导 师

教 授

山东科技大学

二〇二五年五月