E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
©中国图象图形学报版权所有

中图法分类号:TP18;TP391.41　　**文献标识码**:A　　**文章编号**:1006-8961(2026)09-3156-13

论文引用格式:Liu M Y, Xie G S, Shu X B and Zhang L. 2026. Recent advances in open-vocabulary semantic segmentation. Journal of Image and Graphics, 31(9):3156-3168(刘明怡, 谢国森, 舒祥波, 张磊. 2026. 开放词汇语义分割研究进展. 中国图象图形学报, 31(9):3156-3168)[DOI:10.11834/jig.260205]

开放词汇语义分割研究进展

刘明怡[1],谢国森[1*],舒祥波[1],张磊[2]

1. 南京理工大学计算机科学与工程学院,南京　210094; 2. 西北工业大学计算机学院,西安　710129

摘　要:开放词汇语义分割旨在突破传统闭集语义分割对固定类别集合的依赖,使模型能够根据开放类别文本描述实现任意语义目标的像素级识别与分割。以CLIP(contrastive language-image pre-training)为代表的视觉—语言预训练模型虽然为开放类别理解提供了重要支撑,但其侧重于图像级语义对齐,难以满足像素级密集预测对精确定位与细粒度表达的需求。因此,如何实现图像级开放识别能力向像素级精细分割的有效迁移,仍是该领域面临的关键问题。本文对开放词汇语义分割的研究进展进行系统梳理与分析。首先,介绍该任务的研究背景与技术基础,并梳理其与传统语义分割、零样本语义分割之间的演进关系;其次,围绕零样本语义分割、图文监督早期探索、双阶段方法和单阶段方法等主要研究路线,归纳代表性方法的基本思想、技术特点与局限性,并进一步讨论其在遥感图像场景中的拓展应用;再次,总结常用数据集与评价指标,分析现有评测体系的特点与不足;最后,结合当前研究瓶颈,对该领域未来发展方向与研究趋势进行展望。本文可为开放词汇语义分割领域的研究脉络梳理、方法体系比较及后续研究探索提供参考。

关键词:开放词汇语义分割(OVSS);视觉—语言预训练;CLIP;跨类别泛化;开放世界视觉理解

Recent advances in open-vocabulary semantic segmentation

Liu Mingyi[1], Xie Guosen[1*], Shu Xiangbo[1], Zhang Lei[2]

1. School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing　210094, China;
2. School of Computer Science, Northwestern Polytechnical University, Xi'an　710129, China

Abstract: Semantic segmentation is a fundamental task in computer vision that aims to assign a semantic label to each pixel in an image. Most conventional semantic segmentation methods are designed for a closed-set setting in which the categories involved in testing are predefined and largely consistent with those observed during training. Although this setting facilitates the learning of stable category representations and clear decision boundaries, it limits the applicability of segmentation models in real-world scenarios, where unseen objects and novel semantic concepts are frequently encountered. In this context, open-vocabulary semantic segmentation (OVSS) has emerged as an important research direction for extending semantic segmentation from closed-set prediction to open-world visual understanding. OVSS enables models to identify and segment pixel-level regions corresponding to arbitrary textual concepts by using natural language descriptions as category definitions, thereby alleviating the dependence on fixed label spaces. The rapid development of vision-language pre-trained models, especially contrastive language-image pre-training (CLIP) and related large-scale cross-modal models, has provided an important foundation for OVSS. Trained on massive image-text pairs, these models learn aligned visual and tex-

收稿日期:2026-04-16;**修回日期**:2026-05-28;**预印本日期**:2026-06-04

*通信作者:谢国森　guosen.xie@njust.edu.cn

基金项目:国家自然科学基金项目(62276134)

Supported by:National Natural Science Foundation of China(62276134)

tual representations and exhibit strong transferability in open-category recognition. However, most existing vision-language models are mainly optimized for image-level semantic alignment and global visual recognition. Thus, these models have difficulty meeting the requirements of pixel-level dense prediction directly, such as accurate localization, boundary delineation, and fine-grained category discrimination. Therefore, how to effectively transfer image-level open-vocabulary recognition ability to pixel-level segmentation remains a key challenge in OVSS. This paper summarizes recent progress in OVSS from the perspectives of task background, representative methods, benchmark datasets, evaluation metrics, remote sensing extension, and future directions. First, the paper introduces the task background and basic characteristics of OVSS, after which its relationship with conventional semantic segmentation and zero-shot semantic segmentation is clarified. Compared with conventional semantic segmentation, OVSS emphasizes generalization beyond predefined categories, whereas when compared with zero-shot semantic segmentation, it further benefits from large-scale image-text pre-training and allows more flexible category definitions through natural language descriptions. Second, representative OVSS methods are reviewed in accordance with several major research routes, including zero-shot semantic segmentation, early explorations based on image-text supervision, two-stage methods, and single-stage methods. Early zero-shot segmentation methods mainly rely on semantic embeddings to transfer knowledge from seen categories to unseen categories, but their expressive ability is often limited by static semantic representations. Meanwhile, initial studies based on image-text supervision explore whether transferable semantic region representations can be learned without dense pixel annotations, thus providing important inspiration for subsequent OVSS methods. Two-stage methods usually decompose the task into class-agnostic region generation and open-vocabulary region recognition. While this paradigm is clear and modular, its performance is highly dependent on the quality of candidate regions and may incur additional computational cost. In contrast, single-stage methods aim to integrate region modeling, vision-language alignment, and pixel-level prediction within a unified framework, thereby reducing cross-stage error accumulation and improving inference efficiency. Nevertheless, they still face challenges in fine-grained localization, dense semantic alignment, and robust generalization to unseen categories. Apart from analyzing the main ideas, technical characteristics, advantages, and limitations of these methods, this paper discusses the extension of OVSS to remote sensing imagery. Compared with natural images, remote sensing images typically exhibit top-down perspectives, large variations in object scale, significant orientation changes, complex backgrounds, and cross-region distribution shifts, all of which make it difficult to directly transfer OVSS methods designed for natural-image scenarios to remote sensing applications. Therefore, recent studies have begun to explore open-vocabulary remote sensing segmentation frameworks, dedicated benchmark datasets, training-free strategies, and multimodal fusion methods. This development indicates that OVSS is gradually expanding from general natural-image understanding to more complex specialized visual domains. Commonly used datasets and evaluation metrics in OVSS are also summarized in this paper. Existing studies typically use COCO-Stuff as the main training benchmark and evaluate model generalization on datasets, such as Pascal VOC, Pascal Context, and ADE20K, under different vocabulary settings. At present, OVSS evaluation relies on conventional semantic segmentation metrics, especially mean intersection over union (mIoU). However, such metrics emphasize strict category matching and are insufficient for measuring semantic similarity, category hierarchy, and open-category generalization. Therefore, constructing evaluation protocols that better reflect the semantic openness of OVSS remains an important issue. Finally, the paper summarizes the major challenges in current OVSS research and discusses promising future directions, including the effective transfer of open-vocabulary recognition capabilities to precise pixel-level prediction, adaptive region generation, evaluation protocols tailored to open-vocabulary settings, learning with limited supervision, training-free inference, and extensions to domain-specific applications.

Key words: open-vocabulary semantic segmentation (OVSS); vision-language pre-training; contrastive language-image pre-training (CLIP); cross-category generalization; open-world visual understanding

0 引 言

语义分割(semantic segmentation)是计算机视觉(computer vision, CV)中的一项基础任务,其目标是对图像中每个像素进行语义类别标注(Long等,2015;Chen等,2018;Minaee等,2022;严毅 等,2023;高常鑫 等,2024)。长期以来,主流语义分割方法大多建立在固定类别集合的基础之上,即模型训练与测试所涉及的类别基本保持一致。尽管这种闭集(closed-set)设定有助于模型学习稳定的类别表示和清晰的决策边界,但也存在明显局限:当测试场景中出现训练阶段未覆盖的新类别时,模型往往难以做出准确而合理的预测。随着视觉理解任务逐步由封闭环境走向开放环境,传统闭集语义分割方法的适用性正面临越来越大的挑战(Ghiasi等,2022;Sodano等,2024)。

在此背景下,开放词汇学习(open-vocabulary learning, OVL)(Wu等,2024)逐渐成为视觉智能领域的重要研究方向。与传统零样本学习(zero-shot learning, ZSL)(Xian等,2019b)相比,开放词汇学习更加强调自然语言在视觉任务中的建模作用。它不再局限于依赖静态语义嵌入实现类别迁移,而是借助大规模图文对数据学习视觉与语言之间更加丰富、灵活的语义对应关系。尤其是视觉—语言预训练模型(vision-language pre-trained model, VLM)的提出,显著推动了该方向的发展(Du等,2022;张浩宇等,2022;Jia等,2021;Li等,2022)。其中,CLIP(contrastive language-image pre-training)(Radford 等,2021)模型通过图像—文本对比学习(image-text contrastive learning)获得了较强的开放类别识别能力,使得以自然语言描述定义视觉类别成为可能,也为开放词汇语义分割(open-vocabulary semantic segmentation, OVSS)(Zhu和Chen,2024)的研究奠定了重要基础。

然而,图像级视觉—语言模型并不能直接满足像素级分割任务的需求。这是因为图像分类(image classification)更侧重于整体语义表征,而语义分割则要求模型具备更加精细的局部感知与边界刻画能力。换言之,模型不仅需要判断图像中“存在什么”,还需要进一步确定目标“位于何处”以及“边界如何划分”。因此,开放词汇语义分割的核心问题在于,如何将视觉—语言模型在开放类别识别中的优势有效迁移到像素级定位与分割任务中。

对这一问题,研究者提出了多种不同的解决思路。如图1所示,早期工作主要从零样本语义分割(zero-shot semantic segmentation, ZSSS)出发,通过引入文本语义嵌入增强模型对未见类别的识别能力;随着CLIP等模型的广泛应用,相关研究逐步转向以视觉—语言对齐(vision-language alignment)为核心的开放词汇分割框架,并进一步发展出区域提议(region proposal)、掩码适配(mask adaptation)、语义校准(semantic calibration)以及像素匹配(pixel matching)等多种技术路线(Xu等,2022b;Liang等,2023;Xie等,2024)。与此同时,开放词汇分割的研究对象也由自然图像逐步扩展至遥感图像(remote sensing images, RSIs)等专业场景,展现出良好的应用潜力与发展前景(Cao等,2024;Ye等,2025;陶超等,2025;支元杰 等,2026)。

基于上述背景,本文围绕开放词汇语义分割展开综述:对该方向的主要研究思路与方法脉络进行系统梳理;介绍常用的数据集与评价指标;进一步讨论其在遥感图像场景中的任务拓展与应用特点;对该领域未来的发展方向与研究趋势进行分析与展望。

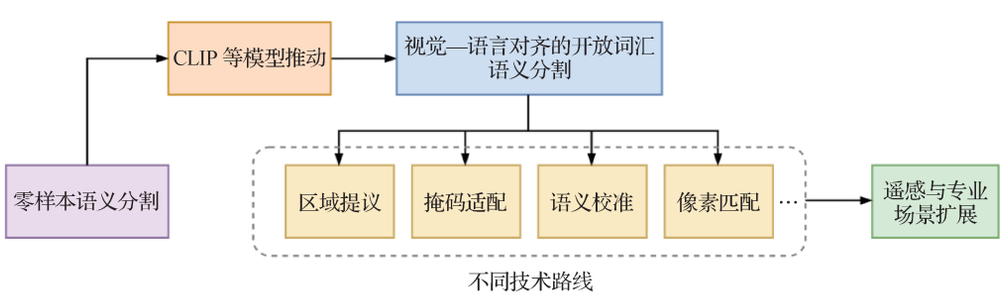


图1 开放词汇语义分割的研究演进与应用拓展示意图

Fig. 1 Research evolution and application expansion of open-vocabulary semantic segmentation

1　方法回顾

开放词汇语义分割的研究演进大致可归纳为零样本分割解耦、基于图文监督的方法、双阶段方法以及单阶段方法等几个主要研究路线。为系统梳理该领域的研究进展，并清晰呈现不同技术路线的基本结构与实现流程，本文对开放词汇语义分割主要方法的典型流程进行了归纳，如图 2 所示。在此基础上，进一步从技术路线、代表性方法、核心思想、主要优势及局限性等维度，对开放词汇语义分割相关研究进行归纳与比较，具体如表 1 所示。与此同时，随着开放词汇分割逐步面向真实应用场景拓展，其研究对象已由自然图像延伸至遥感图像等专业视觉领域，相关代表性工作如表 2 所示。

1.1　从零样本语义分割到开放词汇语义分割

开放词汇语义分割是在零样本语义分割基础上进一步发展而来的。早期零样本语义分割主要围绕已见类别与未见类别之间的语义迁移展开，代表性方法包括 ZS3Net（Bucher 等，2019）、SPNet（Xian 等，2019a）等。这类方法借助类别语义嵌入，通过视觉—语义投影或语义条件下的视觉特征生成，实现对未见类别的像素级预测。

然而，上述方法普遍存在两方面不足：一方面，其所依赖的语义先验相对静态，难以充分表达自然语言中更加丰富和灵活的类别语义；另一方面，其以像素级分类为核心的建模方式与后续兴起的视觉—语言预训练模型之间衔接有限，因而难以有效利用大规模图文预训练所带来的开放语义表征能力。真正推动该方向向开放词汇设定过渡的关键工作是 ZegFormer（Ding 等，2022）。该方法明确指出，将零样本分割直接看做像素级零样本分类并不利于整合图文预训练模型，因此将任务解耦为“类无关区域生成”和“区域级零样本分类”两个环节，为后续开放词汇语义分割方法的发展提供重要思路。

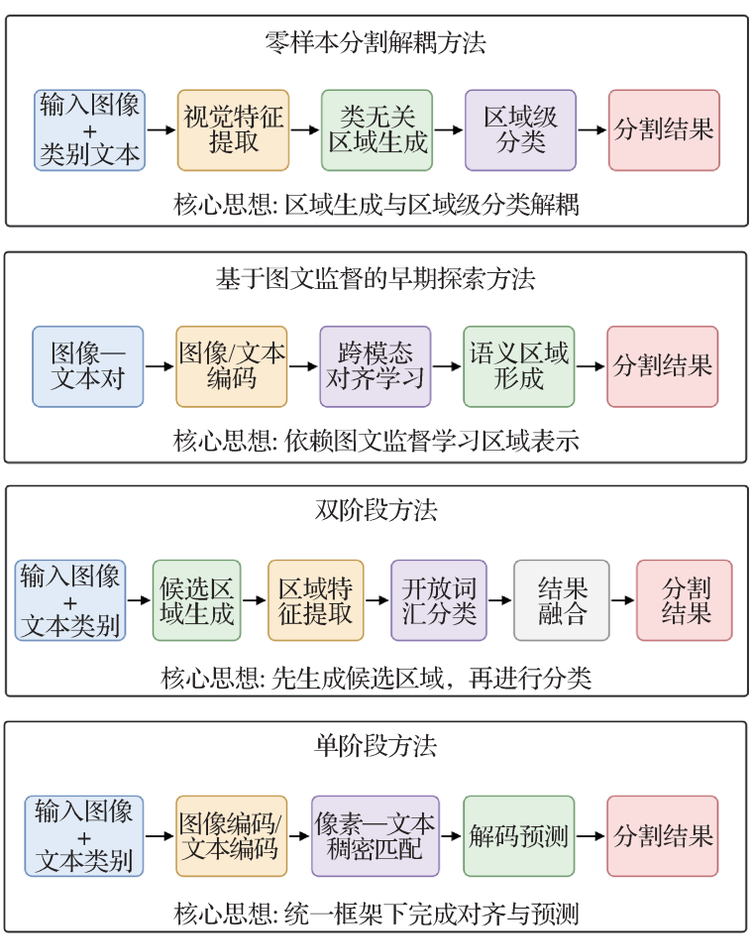


图 2　开放词汇语义分割主要技术路线的典型流程示意图

Fig. 2　Typical workflows of major technical routes for open-vocabulary semantic segmentation

表1　开放词汇语义分割主要研究阶段及代表方法

Table 1　Major research stages and representative methods in open-vocabulary semantic segmentation

研究路线	核心内容	代表方法	优势	局限
从零样本语义分割到开放词汇语义分割	关注已见类向未见类的语义迁移	ZS3Net(Bucher等,2019)	为开放词汇语义分割提供任务基础	开放性有限,对复杂类别和真实场景泛化不足
基于图文监督的早期研究	依托图文监督学习可迁移的语义区域表示,减少对像素级标注的依赖	GroupViT(Xu等,2022a)、TCL(Cha等,2023)	弱化对密集标注的需求,提供重要方法启发	空间定位能力和任务适配性仍有限
双阶段方法	先生成候选区域或掩码,再进行开放词汇分类	OVSeg(Liang等,2023)、ODISE(Xu等,2023a)、SCAN(Liu等,2024)、Mask-Adapter(Li等,2025b)	框架清晰,模块化强,便于利用视觉—语言模型	推理开销较大,易受候选区域质量影响
单阶段方法	在统一框架内完成区域建模、语义对齐和像素级预测	SAN(Xu等,2023b)、FC-CLIP(Yu等,2023)、CAT-Seg(Cho等,2024)、ESC-Net(Lee等,2025)	推理紧凑,误差传播较少,适合端到端优化	像素级适配和开放类别泛化仍具挑战

表2　开放词汇遥感图像语义分割主要研究进展

Table 2　Recent advances in open-vocabulary remote sensing image semantic segmentation

研究阶段	核心内容	代表方法
任务引入与初步探索	面向遥感场景构建开放词汇分割框架,并针对方向变化与尺度变化进行特征建模	OVRS(Cao等,2024)
任务定义与数据集构建	对开放词汇遥感图像语义分割任务进行形式化定义,并提供专用数据集支撑	GSNet(Ye等,2025)
免训练开放分割	将免训练开放词汇分割方法引入遥感场景,并验证其可行性	SegEarth-OV(Li等,2025a)
多模态开放词汇分割	面向光学影像与SAR数据开展跨模态表征对齐与双编码器特征融合	MM-OVSeg(Wei等,2026)

与传统零样本语义分割相比,开放词汇语义分割在任务设定上具有更强的开放性。该任务不仅要求模型具备对未见类别的泛化能力,还允许直接引入文本词表、图像描述及其他语言监督信号,从而在更大范围内实现视觉语义与语言语义的对齐。因此,开放词汇语义分割并非零样本语义分割的简单延伸,而是在视觉—语言预训练快速发展背景下形成的一类更具通用性和应用价值的密集预测任务。

1.2　基于图文监督的开放词汇语义分割早期研究

在开放词汇语义分割研究的早期阶段,部分工作首先关注一个基础性问题,即在仅依赖图文监督而缺乏像素级标注的条件下,模型能否学习到具有可迁移性的语义区域表示。GroupViT(grouping vision Transformer)(Xu等,2022a)是其中较具代表性的工作。该方法在Vision Transformer中引入分层分组机制,使图像块在网络内部逐步聚合为语义区域,并在仅利用成对图文数据进行对比学习的条件下实现零样本语义分割,表明语义区域的形成并不必然依赖显式的像素级监督。TCL(text-grounded contrastive learning)(Cha等,2023)则针对早期图文对比学习方法中训练阶段侧重图像—文本对齐、测试阶段却要求区域—文本匹配的训练—测试不一致问题,提出文本引导的区域—文本对齐机制,使模型能够围绕给定文本生成对应的分割结果,推动开放世界分割由粗粒度图文对应向细粒度区域语义建模发展。进一步地,SegCLIP(Luo等,2023)从视觉表示学习的角度出发,通过带有可学习中心的图像块聚合机制将离散图像块组织为语义区域,并结合重建约束与伪标签约束增强区域一致性,在无掩码标注条件下实现开放词汇分割。

总体而言,这些工作共同说明,即使在缺乏密集像素标注的条件下,模型仍有可能依托图文监督学习到具有空间结构的语义表示,为后续开放词汇语义分割研究提供了重要铺垫。

1.3 开放词汇语义分割的双阶段方法

双阶段方法是开放词汇语义分割中较早形成体系的一类代表性技术路线。其基本思路是将任务拆解为两个相对独立但相互衔接的子问题：首先生成类别无关的候选掩码或候选区域，其次利用视觉—语言模型对这些候选区域进行开放词汇条件下的语义判别。

从研究脉络来看，ZSSeg（a simple baseline for open-vocabulary semantic segmentation）（Xu 等，2022b）可视为双阶段方法的代表性起点。该工作系统地建立了以“候选掩码生成—区域级开放分类”为核心的双阶段框架：第 1 阶段提取具有较强泛化能力的候选掩码，第 2 阶段利用 CLIP 模型对掩码裁剪区域进行分类。随着这一范式的发展，研究者逐渐认识到，制约该类方法性能的主要因素并不完全在于候选掩码生成质量，更关键的问题在于原始视觉—语言模型对经过掩码裁切后的区域图像缺乏足够的适应能力。OVSeg（open-vocabulary semantic segmentation with mask-adapted CLIP）（Liang 等，2023）正是在这一背景下提出的代表性工作。该方法通过构造“掩码区域—文本描述”训练对，对视觉—语言模型进行面向掩码输入的适配，并引入掩码提示调优（mask prompt tuning）机制，在尽量保留上下文信息的同时增强区域级语义识别能力。相较于早期方法将 CLIP 直接作为区域分类器加以使用，OVSeg 进一步表明，双阶段框架的关键并不只是“是否采用掩码提议”，更在于“如何使视觉—语言模型更好地适应候选区域输入”。

在此基础上，双阶段方法的研究重点开始由“候选区域生成”转向“候选区域表征校准”。SCAN（semantic-assisted calibration network）（Liu 等，2024）代表了这一阶段的重要进展。该工作指出，现有开放词汇语义分割方法普遍采用“分割模型预测—CLIP 附加分类”的框架，但该策略容易导致两方面问题：1）训练过程中候选区域嵌入容易向已见类别塌缩，从而削弱模型对未见类别的泛化能力；2）CLIP 在区域级输入条件下易受到领域偏置和上下文缺失的共同影响，进而导致语义判别不稳定。针对上述问题，SCAN 在候选区域特征中注入由 CLIP 提供的广义语义先验，并通过语义辅助校准与上下文偏移修正机制重建候选区域表征与文本语义之间的对应关系。与早期双阶段方法相比，SCAN 更加强调候选区域嵌入空间的显式校准与语义纠偏，推动双阶段方法由粗粒度区域分类向更精细的区域语义对齐发展。

此外，双阶段思想还进一步拓展到开放词汇全景分割任务中，其中具有代表性的工作是 ODISE（open-vocabulary diffusion-based panoptic segmentation）（Xu 等，2023a）。从方法结构上看，ODISE 延续了双阶段思想，即先基于扩散模型的视觉特征生成候选掩码，随后将区域表征与文本嵌入进行匹配，并融合 CLIP 分支的预测结果，对候选掩码进行开放词汇分类。与早期方法主要依赖 CLIP 等判别式视觉—语言模型不同，ODISE 将扩散模型的中间表征用于密集特征提取与掩码生成，并结合 CLIP 的开放词汇识别能力，实现生成式与判别式模型的协同。尽管其任务边界已超出语义分割范畴，但其方法设计仍可视为双阶段范式在开放词汇分割方向上的重要延伸。

近年来，双阶段方法的优化进一步聚焦于两个更为具体的瓶颈。1）候选区域覆盖不充分：若第 1 阶段未能生成与文本查询相匹配的目标区域，则后续分类过程便难以进行有效补救。针对这一问题，PMP（prompt-guided mask proposal）（Li 等，2024）将文本提示直接引入候选掩码生成过程，通过文本标记与查询标记之间的跨注意力机制生成查询感知的候选掩码，推动第 1 阶段由传统的“类别无关掩码提议”向“弱文本感知掩码提议”转变。2）掩码池化带来的表征失真：Mask-Adapter（Li 等，2025b）指出，即使候选掩码本身较为准确，直接在掩码区域内对视觉特征进行池化，也未必能够形成稳定且具有良好可分性的开放词汇语义表征。为此，该方法通过引入语义激活图和掩码一致性损失，增强掩码区域与视觉—语言表征之间的一致性，提高基于掩码池化的双阶段方法在区域级语义判别中的稳定性与鲁棒性。

总体而言，双阶段方法通过任务分解在一定程度上降低了开放词汇语义分割的实现难度，但其性能仍在较大程度上受候选区域质量的制约，且逐区域推理所带来的计算开销相对较高。同时，视觉—语言模型对局部区域输入的适应能力仍有待进一步提升。基于此，后续研究一方面持续围绕区域—语言对齐机制与区域表征能力展开优化；另一方面逐步转向更加高效的端到端单阶段框架。

1.4 开放词汇语义分割的单阶段方法

与双阶段方法通过“候选区域生成—区域级开放词汇分类”流程完成分割不同，单阶段方法更强调在统一框架内同时完成视觉—文本语义对齐与像素级预测。其核心目标在于减少重复特征提取以及跨阶段误差累积，并通过端到端训练或统一推理机制，使图像级视觉—语言预训练表征能够更自然地迁移到像素级预测任务中。总体来看，单阶段方法的发展经历了由预训练视觉—语言模型适配、到像素—文本稠密匹配建模、再到基础模型协同与高效微调的演进过程。

在围绕预训练视觉—语言模型的一体化适配方面，SAN（side adapter network）（Xu等，2023b）是较具代表性的工作。该方法在保持CLIP模型冻结的同时，在其旁路引入轻量级适配网络，并通过掩码预测分支与注意力偏置分支分别生成候选掩码和注意力偏置，引导CLIP完成候选区域的类别识别，使区域预测过程能够更好地对齐CLIP的语义空间。相较之下，FC-CLIP（frozen convolutional CLIP）（Yu等，2023）更明确地将开放词汇语义分割建模为共享主干的单阶段框架，即通过冻结的卷积式CLIP实现特征共享，并在统一框架内完成掩码预测与开放词汇识别，避免对额外候选区域模块的依赖。EBSeg（open-vocabulary semantic segmentation with image embedding balancing）（Shan等，2024）则进一步关注已见类别与未见类别之间的语义平衡问题，通过自适应平衡解码器和语义结构一致性约束增强模型的开放类别泛化能力。该阶段的研究重心已由早期区域语义表征的形成，逐步转向在统一推理框架下对CLIP的高效适配，以及模型开放类别识别能力的提升。

在此基础上，单阶段方法进一步发展为以像素—文本稠密匹配为核心的编码器—解码器范式。CAT-Seg（cost aggregation for open-vocabulary semantic segmentation）（Cho等，2024）将开放词汇语义分割显式建模为图像特征与文本特征之间的代价聚合问题，通过空间聚合与类别聚合增强局部匹配能力和全局语义一致性。SED（a simple encoder-decoder for open-vocabulary semantic segmentation）（Xie等，2024）则提出了一种更为简洁的编码器—解码器结构，其首先利用层次化编码器建立像素与文本之间的代价表示，随后通过逐层融合的解码器恢复空间细节，并借助类别提前筛除机制降低无关类别带来的计算冗余。与SAN、FC-CLIP等更侧重视觉—语言模型适配的方法相比，这类工作将研究重点推进到更稳定的像素级匹配建模层面，标志着单阶段路线由共享特征的一体化预测进一步演进为基于代价图的统一像素级建模。

进一步地，单阶段方法又朝着基础模型协同建模与参数高效适配方向发展。ESC-Net（effective SAM combination network）（Lee等，2025）利用SAM（segment anything model）解码器的类别无关分割能力，并将由图文相关性生成的伪提示嵌入统一推理框架，实现更精细的空间聚合与掩码预测。Hyper-CLIP（Peng等，2025）则从表示空间几何结构出发，通过双曲空间中的缩放变换对CLIP进行参数高效微调，以改善图像级表征向像素级表征迁移过程中的层级对齐。此外，Dual Semantic Guidance（Wang等，2025）通过联合利用图像与文本两侧的语义信息生成像素级伪标注，增强模型的细粒度识别能力；S-Seg（simple open-vocabulary semantic segmentation）（Lai，2025）则进一步表明，开放词汇语义分割可通过伪掩码与语言监督直接训练标准分割架构，在保持框架简洁的同时获得较强的泛化性能。

单阶段方法的进一步演进还体现在局部—全局协同建模与新型视觉主干迁移方面。LoGoSeg（Chen等，2026）通过对象存在先验、区域感知对齐模块与双流融合机制，将全局图文相似度、局部区域约束和全局语义上下文统一到同一模型之中。dinov3. seg（Dutta等，2026）则尝试将DINOv3系列表征引入开放词汇语义分割，通过联合利用全局标记与局部图像块特征，并在图文交互前后进行细化，提升复杂场景中的空间精度与鲁棒性。

总体而言，单阶段方法的研究重心已经由“如何适配CLIP”进一步拓展到“如何整合更强的局部—全局视觉先验并实现统一的像素级视觉—文本建模”。相较于双阶段方法，单阶段路线在提升推理效率和促进端到端协同优化方面展现出一定潜力，但其核心挑战仍是如何将原本面向图像级理解的视觉—语言预训练模型稳定迁移至像素级预测任务，并具备可靠的细粒度空间定位能力与开放类别泛化能力。

1.5 开放词汇遥感图像分割

随着开放词汇分割研究逐步走向真实应用场

景，遥感图像已成为其重要的扩展方向。相较于自然图像，遥感图像通常具有俯视视角、目标方向变化大、尺度跨度大以及跨区域分布差异显著等特点，因此自然图像中的开放词汇语义分割方法难以直接迁移。基于此，OVRS（open-vocabulary remote sensing）（Cao 等，2024）首次较系统地提出了面向遥感图像的开放词汇语义分割框架，并针对方向变化与尺度变化设计了相应的特征建模策略。

随后，GSNet（generalist and specialist network）（Ye 等，2025）进一步对开放词汇遥感图像语义分割（open-vocabulary remote sensing image semantic segmentation，OVRSISS）任务进行了形式化定义，并提出 LandDiscover50K 数据集，为该方向的研究提供了数据支撑。此后，SegEarth-OV（Li 等，2025a）将免训练方法进一步引入遥感场景，在多个遥感数据集和多类任务上验证了无额外训练开放分割方法的可行性。进一步地，MM-OVSeg（multimodal optical-SAR fusion for open-vocabulary segmentation in remote sensing）（Wei 等，2026）又将该方向拓展至光学影像与 SAR（synthetic aperture radar）数据的多模态融合场景，通过跨模态表征对齐与双编码器特征融合，提升了复杂环境下分割的鲁棒性。

总体来看，开放词汇遥感图像语义分割已初步形成从任务定义、数据集构建到方法设计的发展脉络。这表明，开放词汇语义分割的研究对象正在由自然图像场景逐步拓展至专业视觉应用领域，并逐渐发展为具有更强通用性的研究框架。

2 数据集与评价指标

2.1 常用数据集

Pascal VOC（pattern analysis，statistical modeling and computational learning visual object classes）（Everingham 等，2015）是语义分割研究中应用最为广泛的基础数据集之一，包含 20 个前景类别和 1 个背景类，通常划分为 1 464 幅训练图像、1 449 幅验证图像和 1 456 幅测试图像。在开放词汇语义分割任务中，该数据集通常以 PAS-20 的形式使用，主要用于检验模型对典型目标类别的开放识别与分割能力。但由于类别集合相对有限，其在复杂背景语义及更丰富类别场景下的评估能力仍存在一定不足。

COCO-Stuff（common objects in context-stuff）（Caesar 等，2018）是当前开放词汇语义分割研究中最常用的训练基准之一。该数据集在 COCO 原有对象类别标注的基础上，进一步补充了大规模背景区域标注，共包含 172 个类别，其中包括 80 个前景（thing）类、91 个背景（stuff）类和 1 个未标注类，覆盖 COCO 2017（Lin 等，2014）的全部 164 K 幅图像。相较于 Pascal VOC，COCO-Stuff 具有更复杂的场景构成和更丰富的语义层次，整体语义分布也更接近真实世界场景。因此，其被广泛用做开放词汇语义分割的统一训练集，以支撑模型学习更丰富的视觉语义表示。

ADE20K（ADE20K scene parsing dataset）（Zhou 等，2017）是场景理解领域中具有代表性的复杂数据集之一，包含 20 210 幅训练图像、2 000 幅验证图像和 3 352 幅测试图像。在开放词汇语义分割相关研究中，该数据集通常采用两种配置，即 A-150 和 A-847。其中，A-150 对应 150 个常用类别，主要用于评估模型在标准语义分割设定下的整体性能；A-847 对应 847 个更完整的类别集合，更能够反映模型在复杂语义类别条件下的开放泛化能力。

Pascal Context（Mottaghi 等，2014）是对 Pascal VOC 的进一步扩展，提供了更为丰富的场景上下文标注信息，通常划分为约 5 K 幅训练图像和 5 K 幅验证图像。在开放词汇语义分割任务中，该数据集存在 PC-59 和 PC-459 两种常见配置，其中 PC-59 适用于标准化基准对比，PC-459 则更适合考察模型在更大规模类别空间中的泛化能力。

总体而言，上述 4 类数据集已构成开放词汇语义分割中较为稳定的评测体系，如表 3 所示。现有工作通常以 COCO-Stuff 作为训练基准，并在 Pascal VOC（PAS-20）、Pascal Context（PC-59、PC-459）以及 ADE20K（A-150、A-847）上进行测试。该设计不仅能够避免针对目标数据集的特化适配，更客观地衡量模型在开放场景中的泛化能力，同时也便于比较不同方法在中小规模词表与大规模词表条件下的性能差异。

2.2 评价指标

在评价指标方面，开放词汇语义分割目前仍主要沿用传统语义分割中的核心指标平均交并比（mean intersection over union，mIoU）（Zhu 和 Chen，2024）。

然而，需要指出的是，现有语义分割评价指标大

表3　开放词汇语义分割常用数据集

Table 3　Benchmark datasets for open-vocabulary semantic segmentation

数据集	划分	常用配置	主要特点
Pascal VOC(Everingham等,2015)	训练图像1 464幅、验证图像1 449幅、测试图像1 456幅	PAS-20	用于评估典型目标类别的开放识别与分割能力;但类别规模较小
COCO-Stuff(Caesar等,2018)	约164 K幅图像	COCO-Stuff	常用训练基准,场景复杂、语义丰富
ADE20K(Zhou等,2017)	训练图像20 210幅、验证图像2 000幅、测试图像3 352幅	A-150、A-847	A-150用于标准性能评估,A-847用于复杂类别条件下开放泛化评估
Pascal Context(Mottaghi等,2014)	约5 K幅训练/5 K幅验证图像	PC-59、PC-459	PC-59用于标准化对比,PC-459用于大规模类别空间下的泛化评估

多是在闭集设定下形成的,直接用于开放词汇分割任务时存在一定局限。其根本原因在于,闭集评价通常要求预测类别与真实类别在标签层面完全一致,而开放词汇分割更强调预测结果在语义层面的匹配合理性。例如,当模型将“sofa”预测为“couch”时,尽管二者在语义上高度接近,但在严格的类别一致性评价规则下,该预测仍会被判定为错误。基于此,近年来已有研究开始探索更适用于开放词汇分割的评价指标,试图将语义相近类别之间的合理替代关系纳入评价体系之中(Zhou等,2025)。

3　未来研究展望

从现有研究的发展脉络来看,开放词汇语义分割虽然已取得显著进展,但距离构建真正稳定、通用且具备良好泛化能力的开放场景分割系统仍有较大提升空间(Šarić等,2025)。结合前文对不同技术路线的分析,未来该方向仍有若干关键问题值得深入研究。

1)如何实现由“开放类别识别”向“精确像素分割”的有效转化,仍是开放词汇语义分割研究的核心问题。尽管现有方法已表明,视觉—语言预训练模型能够为未见类别提供一定的开放识别能力,但在实际分割过程中,模型在边界定位、小目标解析、细长结构保持以及相邻类别判别等方面仍表现出不足。由此可见,未来研究需要在保持开放语义泛化能力的同时,进一步加强对局部细节、空间结构和区域一致性的建模,从而提升开放词汇分割的整体精度与实用性。

2)区域生成机制仍有进一步优化的空间。如前文对双阶段方法的分析所示,该类方法通常先生成候选区域,再进行开放词汇分类,其性能在较大程度上依赖第1阶段候选掩码的质量。现有方法虽然能生成一定质量的候选区域,但容易受训练数据分布和类别先验影响,存在训练偏置,难以适应测试阶段开放类别的动态变化。因此,未来研究需要进一步探索开放类别感知的自适应区域生成机制,使候选区域生成过程能够结合文本类别描述进行动态调整,并与类别判别过程形成紧密协同,缓解候选掩码质量不足带来的误差传播,提升模型在未知类别和复杂场景下的分割鲁棒性与泛化能力。

3)评测体系仍需进一步完善。如前文对评价指标的分析所示,开放词汇语义分割目前仍主要沿用闭集语义分割中的评价指标,但其与开放词汇语义分割的任务属性并不完全契合(Liu等,2025)。相较于闭集分割,开放词汇语义分割更加关注预测结果在语义层面的匹配合理性以及模型的开放泛化能力,而非仅关注标签层面的严格对应关系。因此,未来有必要在现有像素重叠精度指标的基础上,进一步综合考虑类别语义相近性、类别层次结构以及跨数据集标注体系差异等因素,建立更符合开放场景需求的评价标准。

4)免训练推理与低监督学习仍是未来值得关注的研究方向。前文对基于图文监督的早期方法的分析表明,图像—文本监督有助于模型学习可迁移的语义区域表示,并在一定程度上减少对密集像素标注的依赖,但其区域定位精度和任务适配能力仍然有限。随着基础模型能力的持续增强,如何在降低标注依赖和训练成本的同时,充分挖掘其开放语义理解与区域感知能力,成为开放词汇语义分割的重要研究方向。未来可从免训练推理、弱监督学习、伪标签构造和参数高效适配等方面展开探索,其中免

训练方法通过直接利用基础模型已有能力减少额外训练成本，在降低模型适配开销和简化应用流程方面展现出较大的研究潜力与应用价值（Kombol 等，2025）。

5）面向专业领域的应用拓展将成为开放词汇语义分割的重要发展方向。如前文对开放词汇遥感图像分割的分析所示，开放词汇语义分割的研究对象已由自然图像逐步拓展至遥感图像等专业领域，并在医学影像、工业检测和自动驾驶等场景中展现出广阔的应用前景（Dahal 等，2025；Sun 等，2025；Reichard 等，2025）。相较于自然图像，这些场景往往面临更高的标注成本、更显著的领域差异以及更复杂的任务需求，因此更能体现开放词汇语义分割的实际意义。随着领域数据、应用需求和模型能力的持续积累，开放词汇语义分割在专业场景中的应用拓展有望成为新的研究热点。

4　结　语

开放词汇语义分割是在视觉—语言预训练快速发展背景下形成的重要研究方向，其核心目标在于突破传统闭集语义分割对固定类别集合的依赖，实现面向任意文本类别的像素级语义理解。相较于传统语义分割与零样本语义分割，开放词汇语义分割在任务设定上具有更强的开放性和更高的实际应用价值，因而近年来受到了广泛关注。

综合来看，开放词汇语义分割的研究已经形成了较为清晰的发展脉络。早期工作主要由零样本语义分割演化而来，研究重点集中于已见类别与未见类别之间的语义迁移问题；随后，随着视觉—语言模型的引入，开放词汇语义分割逐步摆脱对封闭类别集合的依赖，开始围绕图文语义对齐、区域表征学习以及像素级开放类别预测等关键问题展开；进一步地，随着基础模型能力的持续增强，相关研究又在统一建模、高效推理、细粒度定位以及复杂场景泛化等方向不断发展。总体而言，开放词汇语义分割正由“开放类别识别在像素级任务中的扩展”逐步迈向“面向开放场景的通用密集视觉理解”这一更高层次的研究目标。

尽管如此，现有研究仍面临若干关键问题有待解决。例如，如何进一步缩小视觉—语言模型在图像级识别与像素级分割之间的能力差距，如何提升区域生成机制对开放类别定义变化的适应能力，如何构建更符合开放场景需求的评测体系，以及如何在尽可能减少额外训练成本和监督依赖的条件下实现高质量分割，仍然是该领域未来发展的重要方向。此外，遥感图像、医学影像、工业检测和自动驾驶等专业场景的持续拓展，也表明开放词汇语义分割不仅具有方法研究意义，同时具备广阔的实际应用前景。

因此，围绕该方向开展更加系统和深入的研究，不仅有助于推动语义分割任务本身的发展，也将为通用人工智能背景下的视觉理解研究提供新的思路与技术支撑。

参考文献（References）

Bucher M，Vu T H，Cord M and Pérez P. 2019. Zero-shot semantic segmentation//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver，Canada：Curran Associates Inc.：#43［DOI：10.5555/3454287.3454330］

Caesar H，Uijlings J and Ferrari V. 2018. COCO-Stuff：thing and stuff classes in context//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City，USA：IEEE：1209-1218［DOI：10.1109/CVPR.2018.00132］

Cao Q L，Chen Y T，Ma C and Yang X K. 2024. Open-vocabulary remote sensing image semantic segmentation［EB/OL］.［2026-04-15］. https://arxiv.org/pdf/2409.07683.pdf

Cha J，Mun J and Roh B. 2023. Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver，Canada：IEEE：11165-11174［DOI：10.1109/CVPR52729.2023.01074］

Chen J Y，Lyu X B，Kou Z Q，Sheng X D，Xu N and Qiao Y G. 2026. LoGoSeg：integrating local and global features for open-vocabulary semantic segmentation//Proceedings of the 40th AAAI Conference on Artificial Intelligence. Singapore，Singapore：AAAI Press：2886-2894［DOI：10.1609/aaai.v40i4.37279］

Chen L C，Papandreou G，Kokkinos I，Murphy K and Yuille A L. 2018. DeepLab：semantic image segmentation with deep convolutional nets，atrous convolution，and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence，40（4）：834-848［DOI：10.1109/TPAMI.2017.2699184］

Cho S，Shin H，Hong S，Arnab A，Seo P H and Kim S. 2024. CAT-Seg：cost aggregation for open-vocabulary semantic segmentation//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle，USA：IEEE：4113-4123［DOI：10.1109/CVPR52733.2024.00394］

Dahal L, Bhandari Y, Segars W P and Lo J Y. 2025. Five models for five modalities: open-vocabulary segmentation in medical imaging// CVPR 2025 Workshop on Foundation Models for 3D Biomedical Image Segmentation. Nashville, USA: Springer: 51-63 [DOI: 10.1007/978-3-032-23496-4_4]

Ding J, Xue N, Xia G S and Dai D X. 2022. Decoupling zero-shot semantic segmentation//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 11573-11582 [DOI: 10.1109/CVPR52688.2022.01129]

Du Y F, Liu Z K, Li J Y and Zhao W X. 2022. A survey of vision-language pre-trained models//Proceedings of the 31st International Joint Conference on Artificial Intelligence. Vienna, Austria: IJCAI: 5436-5443 [DOI: 10.24963/ijcai.2022/762]

Dutta S, Banerjee B and Rezatofighi H. 2026. dinov3.seg: open-vocabulary semantic segmentation with DINOv3 [EB/OL]. [2026-04-15]. https://arxiv.org/pdf/2603.19531.pdf

Everingham M, Eslami S M A, Van Gool L, Williams C K I, Winn J and Zisserman A. 2015. The PASCAL visual object classes challenge: a retrospective. International Journal of Computer Vision, 111(1): 98-136 [DOI: 10.1007/s11263-014-0733-5]

Gao C X, Xu Z Z, Wu D Y, Yu C Q and Sang N. 2024. Deep learning-based real-time semantic segmentation: a survey. Journal of Image and Graphics, 29(5): 1119-1145 (高常鑫, 徐正泽, 吴东岳, 余昌黔, 桑农. 2024. 深度学习实时语义分割综述. 中国图象图形学报, 29(5): 1119-1145) [DOI: 10.11834/jig.230659]

Ghiasi G, Gu X Y, Cui Y and Lin T Y. 2022. Scaling open-vocabulary image segmentation with image-level labels//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 540-557 [DOI: 10.1007/978-3-031-20059-5_31]

Jia C, Yang Y F, Xia Y, Chen Y T, Parekh Z, Pham H, et al. 2021. Scaling up visual and vision-language representation learning with noisy text supervision//Proceedings of the 38th International Conference on Machine Learning. [s.l.]: PMLR: 4904-4916

Kombol N, Martinović I and Šegvić S. 2025. A survey on training-free open-vocabulary semantic segmentation [EB/OL]. [2026-04-15]. https://arxiv.org/pdf/2505.22209.pdf

Lai Z H. 2025. Exploring simple open-vocabulary semantic segmentation//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 30221-30230 [DOI: 10.1109/CVPR52734.2025.02813]

Lee M, Cho S, Lee J, Yang S, Choi H, Kim I J, et al. 2025. Effective SAM combination for open-vocabulary semantic segmentation//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 26081-26090 [DOI: 10.1109/CVPR52734.2025.02429]

Li J N, Li D X, Xiong C M and Hoi S. 2022. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation//Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR: 12888-12900

Li K Y, Liu R X, Cao X Y, Bai X R, Zhou F, Meng D Y, et al. 2025a. SegEarth-OV: towards training-free open-vocabulary segmentation for remote sensing images//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 10545-10556 [DOI: 10.1109/CVPR52734.2025.00986]

Li Y J, Zhang X Y, Wan K, Yu L T, Kale A and Lu X. 2024. Prompt-guided mask proposal for two-stage open-vocabulary segmentation [EB/OL]. [2026-04-15]. https://arxiv.org/pdf/2412.10292.pdf

Li Y K, Cheng T H, Feng B, Liu W Y and Wang X G. 2025b. Mask-Adapter: the devil is in the masks for open-vocabulary segmentation//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 14998-15008 [DOI: 10.1109/CVPR52734.2025.01397]

Liang F, Wu B C, Dai X L, Li K P, Zhao Y N, Zhang H, et al. 2023. Open-vocabulary semantic segmentation with mask-adapted CLIP// Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 7061-7070 [DOI: 10.1109/CVPR52729.2023.00682]

Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. 2014. Microsoft COCO: common objects in context//Proceedings of the 13th European Conference on Computer Vision. Zurich, Switzerland: Springer: 740-755 [DOI: 10.1007/978-3-319-10602-1_48]

Liu Y, Bai S L, Li G B, Wang Y T and Tang Y S. 2024. Open-vocabulary segmentation with semantic-assisted calibration//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 3491-3500 [DOI: 10.1109/CVPR52733.2024.00335]

Liu Y, Wu S L, Bai S L, Wang J H, Wang Y T and Tang Y S. 2025. Stepping out of similar semantic space for open-vocabulary segmentation//Proceedings of 2025 IEEE/CVF International Conference on Computer Vision. Honolulu, USA: IEEE: 22664-22675 [DOI: 10.1109/ICCV51701.2025.02104]

Long J, Shelhamer E and Darrell T. 2015. Fully convolutional networks for semantic segmentation//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA: IEEE: 3431-3440 [DOI: 10.1109/CVPR.2015.7298965]

Luo H S, Bao J W, Wu Y Z, He X D and Li T R. 2023. SegCLIP: patch aggregation with learnable centers for open-vocabulary semantic segmentation//Proceedings of the 40th International Conference on Machine Learning. Honolulu, USA: JMLR.org: #956

Minaee S, Boykov Y, Porikli F, Plaza A, Kehtarnavaz N and Terzopoulos D. 2022. Image segmentation using deep learning: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(7): 3523-3542 [DOI: 10.1109/TPAMI.2021.3059968]

Mottaghi R, Chen X J, Liu X B, Cho N G, Lee S W, Fidler S, et al.

2014. The role of context for object detection and semantic segmentation in the wild//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 891-898 [DOI: 10.1109/CVPR.2014.119]

Peng Z L, Xu Z Q, Zeng Z L, Wen C S, Huang Y, Yang M L, et al. 2025. Understanding fine-tuning CLIP for open-vocabulary semantic segmentation in hyperbolic space//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 4562-4572 [DOI: 10.1109/CVPR52734.2025.00430]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. [s.l.]: PMLR: 8748-8763

Reichard K, Brasch N, Navab N and Tombari F. 2025. Language-guided open-world anomaly segmentation [EB/OL]. [2026-04-15]. https://arxiv.org/pdf/2512.01427.pdf

Šarić J, Martinović I, Kristan M and Šegvić S. 2025. What holds back open-vocabulary segmentation?//Proceedings of 2025 IEEE/CVF International Conference on Computer Vision Workshops. Honolulu, USA: IEEE: 4315-4325 [DOI: 10.1109/ICCVW69036.2025.00448]

Shan X H, Wu D Y, Zhu G L, Shao Y J, Sang N and Gao C X. 2024. Open-vocabulary semantic segmentation with image embedding balancing//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 28412-28421 [DOI: 10.1109/CVPR52733.2024.02684]

Sodano M, Magistri F, Nunes L, Behley J and Stachniss C. 2024. Open-world semantic segmentation including class similarity//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 3184-3194 [DOI: 10.1109/CVPR52733.2024.00307]

Sun B Y, Liu Y H, Wang X X, Tian B, Chen L and Wang F Y. 2025. 3D annotation-free learning by distilling 2D open-vocabulary segmentation models for autonomous driving//Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI Press: 7078-7086 [DOI: 10.1609/aaai.v39i7.32760]

Tao C, Guo X, Hu K Y, Shen Y X and Wang H. 2025. Language-guided cross-spatiotemporal domain adaptation for remote sensing image semantic segmentation. Journal of Image and Graphics, 30(9): 3153-3170 (陶超, 郭鑫, 胡柯彦, 沈羽翔, 王昊. 2025. 以语言为媒介的遥感图像跨时空领域自适应语义分割. 中国图象图形学报, 30(9): 3153-3170) [DOI: 10.11834/jig.240640]

Wang Z Y, Feng T L, Lyu F, Shang F H, Feng W and Wan L. 2025. Dual semantic guidance for open vocabulary semantic segmentation//Proceedings of 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 20212-20222 [DOI: 10.1109/CVPR52734.2025.01882]

Wei Y M, Xiao A R, Chen H R X, Xia J S and Yokoya N. 2026. MM-OVSeg: multimodal optical-SAR fusion for open-vocabulary segmentation in remote sensing [EB/OL]. [2026-04-15]. https://arxiv.org/pdf/2603.17528.pdf

Wu J Z, Li X T, Xu S L, Yuan H B, Ding H H, Yang Y B, et al. 2024. Towards open vocabulary learning: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(7): 5092-5113 [DOI: 10.1109/TPAMI.2024.3361862]

Xian Y Q, Choudhury S, He Y, Schiele B and Akata Z. 2019a. Semantic projection network for zero- and few-label semantic segmentation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 8248-8257 [DOI: 10.1109/CVPR.2019.00845]

Xian Y Q, Lampert C H, Schiele B and Akata Z. 2019b. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(9): 2251-2265 [DOI: 10.1109/TPAMI.2018.2857768]

Xie B, Cao J L, Xie J, Khan F S and Pang Y W. 2024. SED: a simple encoder-decoder for open-vocabulary semantic segmentation//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 3426-3436 [DOI: 10.1109/CVPR52733.2024.00329]

Xu J R, De Mello S, Liu S F, Byeon W, Breuel T, Kautz J, et al. 2022a. GroupViT: semantic segmentation emerges from text supervision//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 18113-18123 [DOI: 10.1109/CVPR52688.2022.01760]

Xu J R, Liu S F, Vahdat A, Byeon W, Wang X L and De Mello S. 2023a. Open-vocabulary panoptic segmentation with text-to-image diffusion models//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 2955-2966 [DOI: 10.1109/CVPR52729.2023.00289]

Xu M D, Zhang Z, Wei F Y, Hu H and Bai X. 2023b. Side adapter network for open-vocabulary semantic segmentation//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 2945-2954 [DOI: 10.1109/CVPR52729.2023.00288]

Xu M D, Zhang Z, Wei F Y, Lin Y T, Cao Y, Hu H, et al. 2022b. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 736-753 [DOI: 10.1007/978-3-031-19818-2_42]

Yan Y, Deng C, Li L, Zhu L K and Ye B. 2023. Survey of image semantic segmentation methods in the deep learning era. Journal of Image and Graphics, 28(11): 3342-3362 (严毅, 邓超, 李琳, 朱凌坤, 叶彪. 2023. 深度学习背景下的图像语义分割方法综述. 中国图象图形学报, 28(11): 3342-3362) [DOI: 10.11834/jig.220292]

Ye C Y, Zhuge Y Z and Zhang P P. 2025. Towards open-vocabulary

remote sensing image semantic segmentation//Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI Press: 9436-9444 [DOI: 10.1609/aaai.v39i9.33022]

Yu Q H, He J, Deng X Q, Shen X H and Chen L C. 2023. Convolutions die hard: open-vocabulary segmentation with single frozen convolutional CLIP//Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: #1399 [DOI: 10.52202/075280-1399]

Zhang H Y, Wang T B, Li M Z, Zhao Z, Pu S L and Wu F. 2022. Comprehensive review of visual-language-oriented multimodal pre-training methods. Journal of Image and Graphics, 27(9): 2652-2682 (张浩宇, 王天保, 李孟择, 赵洲, 浦世亮, 吴飞. 2022. 视觉语言多模态预训练综述. 中国图象图形学报, 27(9): 2652-2682) [DOI: 10.11834/jig.220173]

Zhi Y J, Jiang Y W, Yang Z, Chen Y Z, Hao W K, Ma M Y, et al. 2026. Development status and prospects of pretrained foundation models for remote sensing imagery. Journal of Image and Graphics, 31(4): 973-986 (支元杰, 姜艺伟, 杨知, 陈奕州, 郝文魁, 马明阳, 等. 2026. 面向遥感图像的预训练基础模型发展现状与展望. 中国图象图形学报, 31(4): 973-986) [DOI: 10.11834/jig.250424]

Zhou B L, Zhao H, Puig X, Fidler S, Barriuso A and Torralba A. 2017. Scene parsing through ADE20K dataset//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 5122-5130 [DOI: 10.1109/CVPR.2017.544]

Zhou H, Qi L, Shen T C, Huang H, Yang X, Li X T, et al. 2025. Rethinking evaluation metrics of open-vocabulary segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(8): 6780-6796 [DOI: 10.1109/TPAMI.2025.3562930]

Zhu C Y and Chen L. 2024. A survey on open-vocabulary detection and segmentation: past, present, and future. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(12): 8954-8975 [DOI: 10.1109/TPAMI.2024.3413013]

作者简介

刘明怡，女，硕士研究生，主要研究方向为计算机视觉、人工智能与图像分割。E-mail：mingyiliu569@gmail.com

谢国森，通信作者，男，教授，主要研究方向为计算机视觉、人工智能、多模态大模型与工业缺陷检测。

E-mail：guosen.xie@njust.edu.cn

舒祥波，男，教授，主要研究方向为计算机视觉、人工智能、多模态大模型与具身智能。E-mail：shuxb@njust.edu.cn

张磊，男，教授，主要研究方向为计算机视觉、模式识别与高光谱图像分析。E-mail：nwpuzhanglei@nwpu.edu.cn