**中图法分类号**:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(XXXX)XX-0001-11

**论文引用格式**:Wang Xinjing, Gao Ying, Zou Yaqi, Zhu Zhengyu, Xu Chunxue, Zhao Qi. XXXX. Semantic Alignment and Locality-Driven Open-Vocabulary Semantic Segmentation. Journal of Image and Graphics, XX(XX):0001-0011(王馨静, 高颖, 邹亚琦, 朱政宇, 徐春雪, 赵琦. XXXX. 语义对齐与局部感知驱动的开放词汇语义分割. 中国图象图形学报, XX(XX):0001-0011)[DOI:10.11834/jig.250540]

# 语义对齐与局部感知驱动的开放词汇语义分割

王馨静,高颖*,邹亚琦,朱政宇,徐春雪,赵琦

青岛科技大学数据科学学院,山东省青岛市　266000

**摘　要**:**目的**　面向具身智能与人形机器人等需在真实环境中即时感知与决策的系统,针对现有视觉—语言模型在像素级分割中常见的定位模糊及尺度、视角敏感性问题,提出一种无需像素级微调的推理范式,旨在在不改变预训练表征的前提下提升开放词汇语义分割的像素级判别质量与工程可用性。**方法**　构建TG-CLIP框架,以冻结的CLIP为特征源,设计两类推理期算子以增强局部与跨尺度语义:一是文本引导的重校准机制,将文本查询作为条件信号对patch级表征进行语义性重构并回投;二是多视角一致性推理,通过重采样与镜像视角的结果融合来放大跨视角一致的预测信号。所有操作均在前向路径上完成,无需额外标注或网络微调。**结果**　在八个公开基准上的评测显示,TG-CLIP的平均mIoU达45.5%,超越多个现有方法,比排名第二的ProxyCLIP(40.1%)高出4.4个百分点。定性对比实验进一步显示,TG-CLIP在复杂背景与细小结构处能更好地保留目标细节并减少误分割与漏分割,与定量结果相互印证。消融与超参数研究表明,两类算子可累积提升性能:文本引导重校准对温度参数表现出良好鲁棒性,多视角一致性推理在尺度与翻转策略上存在可量化的精度—效率权衡,推荐的部署配置在实用性上取得较好折中。**结论**　TG-CLIP在保持视觉—语言模型零样本能力的前提下,通过两类轻量的推理算子实现了对像素级语义一致性与尺度、视角鲁棒性的显著改进,为免训练的开放词汇分割提供了一种易于部署且效果稳健的工程化方案。代码链接:https://www.scidb.cn/s/7baeYf。

**关键词**:开放词汇分割;零样本学习;视觉—语言模型;CLIP;像素级语义分割;免训练推理

# Semantic Alignment and Locality-Driven Open-Vocabulary Semantic Segmentation

Wang Xinjing, Gao Ying, Zou Yaqi, Zhu Zhengyu, Xu Chunxue, Zhao Qi

*Qingdao University of Science and Technology, Qingdao, Shandong Province 266000, China*

**Abstract**: **Objective** For embodied intelligence and humanoid robotic systems that demand real-time and robust perception, accurate and fine-grained scene understanding remains a key challenge. In particular, when applying vision-language models to pixel-level segmentation, issues such as localization blur and sensitivity to scale and viewpoint often hinder reliable spatial reasoning. Open-vocabulary semantic segmentation (OVSS) offers a promising solution by enabling pixel-level recognition of arbitrary textual categories without relying on exhaustive pixel-wise annotations. However, large-

**收稿日期**:2025-10-29;**修回日期**:2026-01-16

*通信作者:高颖,通信作者,女,副教授,主要研究方向为计算机视觉、深度学习。E-mail:gaoying@qust.edu.cn;高颖 gaoying@qust.edu.cn

**基金项目**:国家自然科学基金(项目编号:62401310);山东省高等学校青创科技支持计划(项目编号:2024KJG053);青岛市自然科学基金(项目编号:25-1-1-121-zyyd-jch)

**Supported by**: National Natural Science Foundation of China (Grant No. 62401310); Shandong Province University Youth Innovation Technology Support Program (Grant No. 2024KJG053); Qingdao Natural Science Foundation (Grant No. 25-1-1-121-zyyd-jch)

scale vision-language models (VLMs) like CLIP are primarily optimized for image- or region-level tasks, leading to limited spatial granularity and degraded dense prediction performance. To address these limitations, this paper proposes a training-free inference paradigm that enhances the pixel-wise discriminability and engineering readiness of VLM-based OVSS frameworks without modifying pretrained model weights, thereby improving their applicability to real-time robotic perception and decision-making. **Method** We introduce TG-CLIP, a practical framework built on a frozen CLIP (ViT-B/16) backbone that implements two lightweight, complementary inference-stage operators designed first to inject text-conditioned semantic bias into local visual representations and second to amplify prediction robustness across scales and viewpoints. The first operator, Text-Guided Recalibration (TGR), treats text queries as conditional semantic probes: given a set of class queries encoded into normalized query vectors, TGR computes per-location query-visual affinities and back-projects the resulting semantic signal into the local feature space. The injected signal is blended with the original visual features under a tunable injection strength and renormalized so that the final logits retain a cosine-similarity interpretation. Importantly, TGR is a parameter-free forward operator requiring no pixel-level supervision or fine-tuning; it leverages only the pretrained visual and textual encoders to effect a semantic recalibration of local responses. The second operator, Multi-view Consistency Reasoning (MCR), aggregates evidence across a compact set of resampling scales and mirrored views. MCR first resamples the input at multiple scales that are rounded to integer multiples of the model's patch grid to avoid misalignment artifacts, then performs forward inference (optionally via an overlap-aware sliding window) and up-samples logits back to the original image resolution. For each scale, MCR also computes a horizontally mirrored view, maps its outputs back to the canonical coordinate frame, and aggregates the mirrored and canonical predictions. A final cross-scale averaging yields the comprehensive dense prediction. The overall design explicitly preserves the zero-shot characteristics of the underlying VLM while using only forward-time operations to improve spatial precision. **Result** We evaluate TG-CLIP on eight public segmentation benchmarks that represent diverse settings with and without explicit background classes. Benchmarks include PASCAL VOC, PASCAL-Context, COCO-Object, ADE20K, COCO-Stuff, Cityscapes, and commonly used VOC subsets. For feature extraction, we use the CLIP ViT-B/16 encoder in a frozen configuration. All experiments are performed without any additional training or pixel-level fine-tuning; hyperparameters are kept consistent across datasets. Across the evaluated benchmarks, TG-CLIP achieves an average mIoU of 45.5%, outperforming several recent baselines. Notably, TG-CLIP exceeds the second-ranked ProxyCLIP (40.1%) by 4.4 percentage points in averaged mIoU. Qualitative inspections indicate that TG-CLIP yields crisper boundaries, better preservation of thin structures, and reduced spurious activations in complex scenes. Detailed ablation studies confirm that both operators contribute additively to performance gains: TGR primarily improves local category separability and boundary fidelity, while MCR increases robustness to scale and viewpoint variability. Sensitivity analyses show that TGR is robust to a wide range of temperature and injection-strength settings, and MCR exhibits a measurable accuracy-efficiency tradeoff dictated by the number of scales and whether mirrored views are included; we report and recommend practical operating points that balance latency and accuracy for real-world deployment. TG-CLIP demonstrates that carefully designed inference-stage interventions can substantially close the gap between region-optimized VLM representations and the needs of dense, pixel-level segmentation without retraining. Because all modifications occur at inference time, our framework is particularly attractive for engineering-constrained scenarios such as embedded perception or rapid prototyping where re-training is costly or infeasible. Nevertheless, some limitations remain: the multi-scale and mirrored inference steps introduce extra computation proportional to the number of scales and views; performance on extremely small or heavily occluded instances is still bounded by the spatial granularity of the frozen encoder; and our method does not obviate the potential gains from task-specific supervised fine-tuning when labels are available. Finally, while our experiments include multiple standard benchmarks, further validation on real-world robotic perception tasks and diverse domain shifts would better elucidate generalization properties. **Conclusion** We propose TG-CLIP, a training-free inference methodology that combines Text-Guided Recalibration with multi-scale flipped fusion to enhance pixel-level semantic discrimination and robustness of VLM-based open-vocabulary segmentation. TG-CLIP is simple to implement, preserves the zero-shot strengths of pretrained models, and yields consistent empirical gains across diverse benchmarks. We release implementation details and deployment recommendations to facilitate adoption and further research on inference-centric approaches for dense vision-language tasks. Codes are available at: https://www.scidb.cn/s/

7baeYf.

**Key words**: Open-vocabulary semantic segmentation; Zero-shot learning; Vision-language models; Contrastive Language-Image Pre-training(CLIP); Pixel-level semantic segmentation; Training-free inference

## 0 引　言

随着具身智能与人形机器人逐步走向真实复杂环境，视觉感知系统不仅需要在预定义类别上保持准确性，还需具备对未知概念的理解能力、跨场景迁移能力以及在复杂条件下的稳定判别能力。在实际部署中，系统往往难以依赖高质量像素级标注完成环境建模与决策，这使得如何在弱先验甚至无先验条件下实现可靠的像素级感知成为关键挑战。

在此背景下，依托大规模图文对预训练的视觉—语言模型(Vision-Language Models, VLM)，开放词汇语义分割(Open-Vocabulary Semantic Segmentation, OVSS)逐渐成为研究热点。该任务旨在无需目标类别的像素级标注，仅通过任意文本描述实现像素级语义预测，为开放世界感知提供高度可扩展的解决方案(Li等，2022)。CLIP等模型通过对比学习将视觉与文本映射至共享语义空间，在零样本识别任务中展现出良好的跨类泛化能力(Radford等，2021)，为OVSS奠定了重要基础。

然而，VLM的预训练目标主要聚焦于图像级或区域级对齐，其patch级表示在空间精度与局部判别力方面存在先天不足。在密集预测任务中，这一缺陷常表现为细粒度类别混淆、目标边界模糊以及局部响应不稳定，在复杂结构或遮挡区域尤为明显(Zhou等，2023)。围绕上述问题，现有OVSS方法大体沿着以下几条路径展开探索。一类方法通过增强跨模态语义对齐，提升零样本条件下的语义一致性。例如，在像素、patch或区域层面引入多层次相似度建模与推理校准，以缓解视觉与文本之间的粒度不匹配问题(Luo等，2022；Zhou等，2023)。另一类研究引入结构化掩码或区域聚合策略，通过弱监督或额外结构先验改善分割边界与区域一致性(Ghiasi等，2022；Shin等，2024；Xiang等，2024；Bai等，2024)。此外，也有工作通过多任务协同或跨模型交互统一语义空间，将检测、分割或SAM等模型引入开放词汇框架中，以提升泛化能力与结构感知(Zhang等，2023；Yuan等，2024；Yu等，2024；Wang等，2025)。

尽管上述方法在一定程度上改善了OVSS的像素级表现，但多数仍依赖额外训练、结构改造或弱监督信号，引入了较高的计算开销与数据依赖，这在资源受限或需快速部署的具身系统中并不理想。为此，近年来逐渐出现免训练(training-free)OVSS的研究方向，该类方法在冻结模型参数的前提下，仅通过推理阶段的算子设计或统计校正实现开放分割(Luo等，2023；Liu等，2024；Stegmüller等，2024；Bai等，2025)。免训练范式在可部署性与跨域适应性方面具有显著优势，但其性能仍受制于两个核心瓶颈。一方面，CLIP等VLM提供的patch级表征在细粒度类别与复杂局部结构中判别力不足，易产生语义混淆；另一方面，现有推理策略较少显式建模多尺度与多视角信息，导致跨尺度、跨视角响应不一致，从而引发分割边界断裂与区域漂移(Hajimiri，2024；Kombol，2025)。这些问题在开放场景、长尾类别及高风险决策环境中尤为突出，直接影响具身系统的稳定性与可靠性。

针对上述挑战，本文聚焦于严格免训练条件下的开放词汇语义分割，在不引入额外像素标注或模型微调的前提下，探索如何通过推理阶段增强语义对齐能力与局部空间感知能力。为此，我们提出两个相互补充的推理级模块：多视角一致性推理与文本引导重校准机制。前者通过多尺度采样与翻转推理整合不同视角下的预测响应，以抑制单尺度噪声并提升空间一致性；后者以文本查询作为语义调制信号，对patch级响应进行选择性强化与抑制，从而增强细粒度语义判别能力。如图1所示，所提方法显著提升了边界细节、局部结构保持与语义一致性。我们在多个公开基准上评估了各模块独立与协同性能，并分析了实际部署中的适用性与局限，为免训练OVSS在具身系统中的应用提供了可行路径。

## 1 方　法

### 1.1 方法概述

本文提出一个基于冻结CLIP(Radford等，

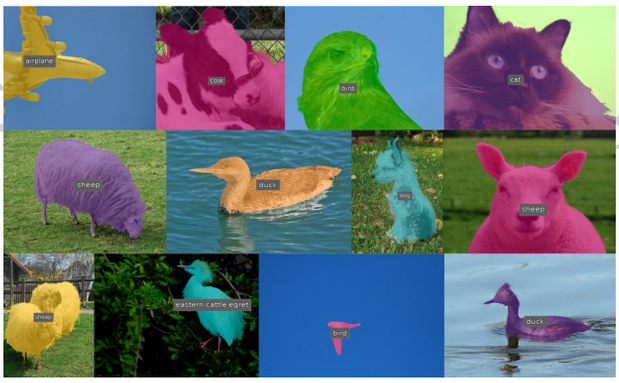



图1　TG-CLIP可视化结果

Fig. 1　TG-CLIP visualization results

2021)的轻量化开放词汇语义分割框架,旨在在无需像素级微调的前提下,通过前向推理阶段的轻量算子显著提升像素级预测质量。不同于依赖额外训练或引入对象级结构的复杂方法,本工作聚焦两个直接、高效且可在推理阶段即插即用的核心模块:多视角一致性推理与文本引导重校准。两者分别提升跨尺度空间一致性与局部语义判别力,使得整个系统在保持零样本泛化的同时,能够获得精细且边界稳定的分割结果。

整个流程遵循单次前向原则:输入图像与类别文本查询首先经CLIP的图像编码器与文本编码器生成多尺度的patch-level视觉特征与文本嵌入。随后,多视角一致性推理模块在每个尺度上并行产生正向视图与镜像视图的预测,镜像预测经映射回原始坐标系后与正向预测在该尺度内完成对齐与融合。通过上采样与配准,模块输出尺度级logits,在尺度之间,则进一步融合来自不同重采样尺度的语义响应,使模型能够整合跨尺度、跨视角的稳定证据。值得一提的是,这里的水平翻转属于推理阶段的数据增强范畴(TTA),但与常规TTA存在两点不同:其一,翻转在每个尺度内部并行执行并在该尺度上完成还原与融合,而不是对整张图的最终输出做简单堆叠或平均;其二,整个过程保持单次前向执行,不进行多轮迭代或参数更新。该设计放大了同时在多个视角和尺度上出现的语义证据,同时抑制仅在某一视图或尺度出现的孤立激活,从而有助于提高区域连通性与边界精度。得到基础空间一致的预测后,文本引导重校准模块以轻量残差的方式将查询语义注入局部patch表征,通过对条件相关特征的动态重构与调制,抑制噪声激活并强化局部语义边界。整体流程如图2所示。

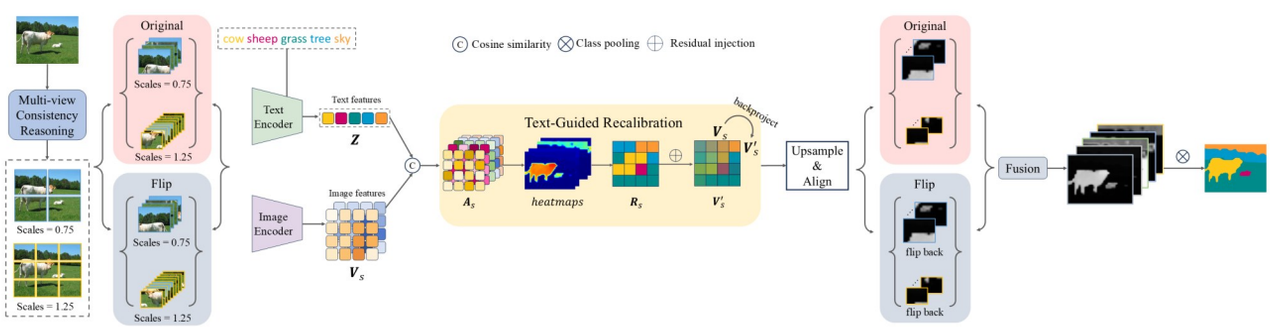



图2　TG-CLIP整体框架图

Fig. 2　TG-CLIP overall framework diagram

## 1.2　多视角一致性推理

为在免训练条件下提升像素级分割的稳健性与边界一致性,本文在推理端引入多视角一致性推理机制。设输入图像为$\boldsymbol{X}\in\boldsymbol{R}^{3\times H\times W}$,文本类别描述集合为$\{t_c\}_{c=1}^{C}$,CLIP patch大小为$p$。文本首先经文本编码器$\mathrm{T}(\cdot)$映射为单位化的查询向量集合$\boldsymbol{Z}=\{z_c\}_{c=1}^{C}$:

$$z_c=\frac{\mathrm{T}(t_c)}{\|\mathrm{T}(t_c)\|},\quad \boldsymbol{Z}\in\boldsymbol{R}^{C\times d} \tag{1}$$

图像端由图像编码器$\varepsilon(\cdot)$产生空间化的patch表征,为了获得真实的多尺度表征并保持与CLIP预训练patch网格的一致性,对输入图像在尺度集合$\boldsymbol{S}=\{s_1,\cdots,s_{|S|}\}$上进行整图缩放。计算每个尺度$s$下的补齐高度$\tilde{H}_s$与宽度$\tilde{W}_s$以保证后续patch对齐:

$$\tilde{H}_s=\left\lceil\frac{sH}{p}\right\rceil p,\quad \tilde{W}_s=\left\lceil\frac{sH}{p}\right\rceil p \tag{2}$$

随后将输入按该分辨率进行重采样,得到重采样图像$\boldsymbol{X}_s$。图像编码器对$\boldsymbol{X}_s$生成稠密视觉特征$\boldsymbol{V}_s=\varepsilon(\boldsymbol{X}_s)$,其中不同尺度$\boldsymbol{V}_s$由于输入图像被缩放而对应不同的有效感受野,产生尺度变化。在查询集条件

下通过相似度映射得到尺度级初始logits：

$$\boldsymbol{L}_s^{(0)} = \mathrm{sim}(\boldsymbol{V}_s, \boldsymbol{Z}) \tag{3}$$

其中$\mathrm{sim}(\cdot,\cdot)$可取余弦相似度或其他相似性量度，且$\boldsymbol{L}_s^{(0)}$的空间分辨率与$\boldsymbol{V}_s$的patch网格一致。为便于像素级判定，初始logits被上采样并对齐回原始图像网格：

$$\hat{\boldsymbol{L}}_s^{(0)} = \mathrm{u}(\boldsymbol{L}_s^{(0)}; H, W) \tag{4}$$

其中$\mathrm{u}(\cdot;\ H, W)$为上采样与对齐算子。为进一步增强局部特征的判别性，我们将上采样后的表征输入至文本引导重校准模块（见1.3节）进行处理，得到增强后的logits，其运算过程记为$\mathrm{R}(\cdot;\ \boldsymbol{Z})$：

$$\bar{\boldsymbol{L}}_s = \mathrm{R}(\hat{\boldsymbol{L}}_s^{(0)}, \boldsymbol{Z}) \tag{5}$$

采用将重校准置于翻转融合之前的设计，能够保证所有视图在语义条件上保持一致性。翻转视图采用与原视图相同的查询条件进行重校准，随后再将两者在像素级进行比较与融合。为获得视角互补信息，对每一尺度执行水平翻转操作$\Phi(\cdot)$，并按与原视图等价的流程得到翻转视图的重校准输出：

$$\bar{\boldsymbol{L}}_s^{flip} = \Phi\left(\bar{\boldsymbol{L}}_s\right) \tag{6}$$

尺度内的视图融合采用对称平均以保证方向不变性与数值稳定性：

$$\tilde{\boldsymbol{L}}_s = \frac{1}{2}\left(\bar{\boldsymbol{L}}_s + \bar{\boldsymbol{L}}_s^{flip}\right) \tag{7}$$

最终在尺度维度上以稳健的均值策略聚合得到综合logits：

$$\boldsymbol{L}^* = \frac{1}{|S|}\sum_{s\in S}\tilde{\boldsymbol{L}}_s \tag{8}$$

对得到的$\boldsymbol{L}^*$通过缩放系数$\gamma$进行温度缩放以调整softmax的尖锐程度，从而控制预测置信度与类别分布的平滑性，得到每像素的类别概率分布$P = \mathrm{softmax}(\gamma \boldsymbol{L}^*)$，最终像素级预测由$\hat{Y} = \mathrm{argmax}_c P_c$导出。经过多尺度与多视图的融合，该模块的输出已具备了良好的空间一致性与鲁棒性，为最终的高质量分割奠定了基础。

### 1.3 文本引导重校准

文本引导重校准旨在通过查询条件化的语义方向与基于置信度与邻域一致性的权重，在推理端放大那些既语义明确又在空间上连贯的patch响应，抑制孤立噪点与定位模糊的激活，从而在不修改模型参数的前提下提升像素级别的语义区分能力。设尺度$s$下经过图像编码器和上采样得到的视觉表征经展平为矩阵$\boldsymbol{V}_s \in \boldsymbol{R}^{N_s \times d}$，其中$N_s = h_s \cdot w_s$表示该尺度下的空间格点数，$d$表示视觉特征维度，文本查询集合表示为矩阵$\boldsymbol{Z} \in \boldsymbol{R}^{C\times d}$，其中$C$为查询数。为保持与余弦相似度度量的一致性，首先对视觉表征与查询向量分别进行行归一化：

$$\tilde{\boldsymbol{V}}_s = \mathrm{Normalize}(\boldsymbol{V}_s),\quad \tilde{\boldsymbol{Z}} = \mathrm{Normalize}(\boldsymbol{Z}) \tag{9}$$

其中$\mathrm{Normalize}(\cdot)$表示逐行除以其$\mathrm{L}_2$范数，使得每一行向量的范数为1。

基于归一化表征，定义查询—视觉交互的注意力权重矩阵：

$$\boldsymbol{A}_s = \mathrm{softmax}(\tau^{-1}\tilde{\boldsymbol{V}}_s\tilde{\boldsymbol{Z}}^{\mathrm{T}}) \in \boldsymbol{R}^{N_s\times C} \tag{10}$$

其中温度参数$\tau > 0$控制注意力的尖锐程度，$\mathrm{softmax}(\cdot)$对每一空间位置沿查询维度归一化，使得每一位置上对各查询的响应和为1。注意力矩阵$\boldsymbol{A}_s$表征了在给patch上各查询的相对响应强度。

利用注意力权重将查询语义回投到视觉特征空间，得到基于查询的语义增强项：

$$\boldsymbol{R}_s = \boldsymbol{A}_s\tilde{\boldsymbol{Z}} \in \boldsymbol{R}^{N_s\times d} \tag{11}$$

该向量沿查询语义方向对patch特征提供条件性“拉动”。若将$\boldsymbol{R}_s$等量注入到所有位置，容易把原本的孤立噪声一并放大；因此我们采用小幅残差注入的策略，将增强项以受控强度线性融合回原始视觉表征并再次归一化，以在不改变原始尺度与整体结构的前提下稳健增强那些在查询空间上已有响应的patch：

$$\boldsymbol{V}_s' = \mathrm{Normalize}(\boldsymbol{V}_s + \lambda\boldsymbol{R}_s) \tag{12}$$

其中$\lambda \geqslant 0$为注入强度系数，用以平衡视觉原始信息与查询导向增强之间的权重选择。该设计使得只有在查询语义方向上已有显著响应的空间位置会被温和放大，而对高不确定性或与查询语义不一致的孤立激活影响有限，从而有助于抑制噪声并提升像素级语义辨识度。小幅残差注入策略本质上是一种轻量级的特征调制，避免了复杂迭代优化带来的计算开销，适合对实时性要求高的部署场景。

最终的文本条件logits由增强后表征与归一化查询的相似度构成：

$$\bar{\boldsymbol{L}}_s = \boldsymbol{V}_s'\tilde{\boldsymbol{Z}}^{\mathrm{T}} \in \boldsymbol{R}^{N_s\times C} \tag{13}$$

并可按空间格点重排为与图像网格对应的稠密logits地图以供后续融合与概率化处理。

## 2 实 验

### 2.1 实现细节

本文在多个公开数据集上对所提出的TG-CLIP方法进行了系统评估，评价范围覆盖两类典型开放词汇语义分割场景：(1)包含背景类的数据集：PASCALVOC（VOC）(Everingham 等，2010)、PASCALContext(Context)(Mottaghi 等，2014)、COCOObject(COCO-Obj)(Lin 等，2014)；(2)不包含背景类的数据集：PASCALVOC20(VOC20)(Everingham 等，2010)、Cityscapes(City)(Cordts 等，2016)、PASCAL-Context59(Context59)(Mottaghi 等，2014)、ADE20K(ADE)(Zhou 等，2017)以及 COCOStuff(COCO-Stf)(Caesar 等，2018)。

所有实验均采用CLIP的ViT-B/16模型(Radford 等，2021)作为视觉—语言特征抽取器。为了适应不同数据集的图像尺度并兼顾推理吞吐，我们将输入图像短边统一缩放为336像素(Cityscapes由于分辨率较高，短边设为560像素)。随后采用滑动窗口推理，窗口大小为224×224，步幅为112，用以覆盖整幅图像并确保局部特征完整性。在所有数据集上，我们均保持统一的推理流程，不进行任何形式的重训练或微调，所有超参数在所有数据集上统一设置；不使用额外的后处理策略，直接在验证集进行前向推理。

### 2.2 主要结果

#### 2.2.1 定量结果

表1汇总了各种零样本语义分割模型的比较结果。本文提出的TG-CLIP方法在所有评估基准上均表现出色，尤其在具有背景类的PASCAL VOC(66.6%)、COCO-Obj(39.4%)和VOC20(86.8%)等数据集上实现了显著的性能提升。此外，TG-CLIP在无背景类数据集PASCAL Context(37.8%)、Cityscapes(42.5%)以及ADE20K(21.0%)中也领先于其他方法，显示出其良好的泛化能力。

**表1 本文方法与其他方法的性能比较**

**Table 1 A performance comparison of the method proposed in this paper with other methods**

| 方法 | VOC | VOC20 | Context | Context59 | COCO-Obj | COCO-Stf | ADE | City | Avg |
|---|---|---|---|---|---|---|---|---|---|
| CLIP(Radford 等，2021) | 20.8 | 49.1 | 9.3 | 11.2 | 8.9 | 5.7 | 3.2 | 6.7 | 14.4 |
| GroupViT(Xu 等., 2022) | 52.3 | 79.7 | 18.7 | 23.4 | 27.5 | 15.3 | 10.4 | 18.5 | 30.7 |
| CLIPSurgery(Li 等, 2023) | 55.2 | 77.5 | 30.3 | 33.4 | 29.7 | 22.2 | 16.1 | 33.1 | 37.2 |
| MaskCLIP(Ding 等, 2023) | 51.4 | 62.9 | 22.5 | 26.2 | 24.9 | 16.9 | 12.3 | 25.6 | 30.3 |
| SCLIP(Wang 等., 2023) | 59.7 | 81.5 | 31.7 | 34.5 | 33.5 | 22.7 | 16.5 | 32.3 | 39.1 |
| ClearCLIP(Lan 等., 2024) | 57.0 | 82.3 | 32.2 | 35.8 | 32.5 | 24.0 | 17.3 | 32.8 | 39.2 |
| GEM(Bousselham 等, 2024) | 58.7 | 81.7 | 32.0 | 35.6 | 32.9 | 23.9 | 16.9 | 32.6 | 39.3 |
| LaVG(Kang 等, 2024) | 62.1 | 82.5 | 31.6 | 34.7 | 34.2 | 23.2 | 15.8 | 26.2 | 38.8 |
| NACLIP(Hajimiri 等, 2024) | 58.9 | 79.7 | 32.2 | 35.2 | 33.2 | 23.3 | 17.4 | 35.5 | 39.4 |
| ProxyCLIP(Lan 等., 2024) | 56.6 | 76.8 | 34.0 | 37.4 | 34.6 | 25.1 | 18.7 | 37.5 | 40.1 |
| **TG-CLIP(Ours)** | **66.6** | **86.8** | **37.8** | **41.5** | **39.4** | **28.2** | **21.0** | **42.5** | **45.5** |

注：加粗字体为每行最优值。

表2给出了各免训练方法在统一评测配置下的计算开销对比(FPS、参数量、单张推理时间与FLOPs)。从表中可以看出，TG-CLIP在效率与精度之间取得了较好的权衡：虽然其并非最快(CLIP基线在FPS最高)，但与最近的强基线相比依然具有明显优势。具体而言，与使用额外支撑网络且参数量显著更大的ProxyCLIP相比，TG-CLIP的FPS提升约9%，参数量由235.4M降至149.6M，FLOPs则减少约49%。尽管CLIP原始骨干等免训练方法在速度上占优，但其分割性能远低于本方法。

需要注意的是，虽然TG-CLIP在推理速度上落后于部分免训练基线，但由于它免去像素级重训练

与微调这一昂贵步骤，因此与全监督或弱监督OVSS方法相比，TG-CLIP在工程化部署与快速迭代场景中更具实用性和部署便利性。

表2　免训练方法的效率对比

Table 2　Comparison of Train-Free Method Efficiency

| 方法 | FPS ↑ | 参数量(M) ↓ | 推理时间 ↓ | FLOPs (G) ↓ |
|---|---|---|---|---|
| CLIP | 15.2 | 149.6 | 0.1 | 35.2 |
| NACLIP | 8.3 | 149.6 | 0.1 | 32.3 |
| ClearCLIP | 5.4 | 149.6 | 0.1 | 15.8 |
| SCLIP | 1.9 | 149.6 | 0.1 | 45.1 |
| ProxyCLIP | 2.2 | 235.4 | 0.4 | 37.7 |
| TG-CLIP(Ours) | 2.4 | 149.6 | 0.4 | 19.1 |

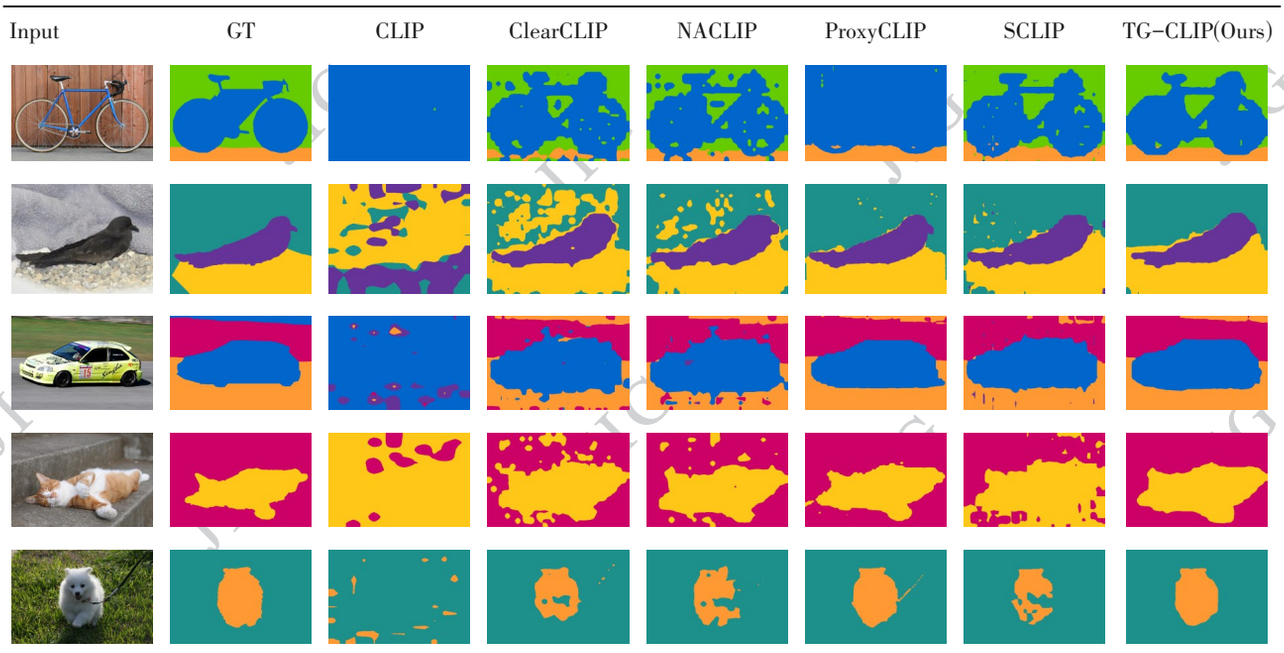
#### 2.2.2　定性结果

图3展示了本方法与多种基线模型在多个典型场景上的定性对比，结果进一步支撑了量化指标中的性能提升。以自行车样例为例，CLIP、ClearCLIP及NACLIP在车架与车轮等结构处出现了明显的背景误激活，导致整体区域破碎或边界模糊。TG-CLIP能够更准确地恢复自行车的主体形状，其预测区域更为连贯、边缘更加清晰。这一改善主要源于文本引导重校准机制：通过查询条件化的注意力回投，TGR将文本语义方向性地注入patch表征，并以轻量残差的方式强化与查询语义一致的空间响应，同时抑制语义不一致的孤立激活，从而有效提升了类别一致性与边界连通性。

在猫和狗等动物样例中，多数基线方法均受到背景纹理干扰，呈现出大面积的噪声块或局部缺失；而TG-CLIP的预测更为稳定，主体区域保持完整，特别是在轮廓边缘与身体细节处表现更佳。这部分优势来自多视角一致性推理机制：MCR将多尺度重采样与镜像视角下的预测进行汇集，并在对齐后执行一致性融合，使跨视角持续出现的真实结构得到增强，而仅出现在单一尺度或单一视角的噪声被有效弱化。因此，TG-CLIP在细小结构、长条状区域及小目标场景中展现出更强的鲁棒性。

图3　本文方法与其他方法的定性对比结果

Fig. 3　Qualitative comparison results of this method with other methods




### 2.3　消融实验

在本节中，我们进行消融实验以验证各个模块对模型性能的贡献。我们通过逐步加入不同的模块来分析每个组件对模型整体效果的提升情况，并为每个模块的有效性提供量化证明。此外，针对模型中的关键组件，我们还进行了调参实验，以进一步优化其性能。以下是对各个模块及其调参结果的详细分析。

#### 2.3.1　各组件的有效性验证

表3展示了逐步合并不同模块后在各数据集上

的性能对比。首先，基于基准模型（baseline），本文实现了在不同数据集上的初步分割性能。接下来，通过加入多视角一致性推理，模型的性能得到了显著提升，尤其在PASCAL VOC（+2.5%）和Cityscapes（+0.8%）等数据集上表现出明显的改进。多视角一致性推理通过增加了多尺度推理和翻转推理的结合，提高了模型在处理复杂背景和多样化视角下的鲁棒性，整体平均mIoU提升了1.25%。

随后引入了文本引导重校准模块。在此基础上，模型的性能再次得到提升，尤其在Cityscapes（+0.7%）和COCO-Stuff（+0.7%）数据集上的表现有所提高。文本引导重校准通过有效加强了局部语义一致性，使得模型在细节捕捉和边界处理上表现更加精准，进一步减少了误分割和漏分割现象。加入该模块后，整体模型的平均mIoU达到了45.5%，相比基准模型提升了1.6%。

**表3　各组件的消融实验**

**Table 3　Ablation experiments of each component**

| 方法 | VOC | VOC20 | Context | Context59 | COCO-Obj | COCO-Stf | ADE | City | Avg |
|---|---|---|---|---|---|---|---|---|---|
| baseline | 64.6 | 84.3 | 36.8 | 40.1 | 37.7 | 26.6 | 20.1 | 41.0 | 43.9 |
| +多视角一致性推理 | 66.5 | 86.8 | 37.7 | 41.2 | 39.1 | 27.5 | 20.6 | 41.8 | 45.1 |
| **+文本引导重校准** | **66.6** | **86.8** | **37.8** | **41.5** | **39.4** | **28.2** | **21.0** | **42.5** | **45.5** |

注：加粗字体为每行最优值。

通过这些逐步实验，我们可以看到每个模块对模型的性能均有显著贡献，提升了模型在复杂场景和细节处理上的能力，验证了它们在提升模型分割精度方面的有效性。

2.3.2　多视角一致性推理的参数调优

为明确多视角一致性推理中尺度集合对性能与推理效率的影响，本文对Scales参数进行了定量调优，如表4所示，实验比较了单尺度与两/三尺度组合的表现，并同时记录平均mIoU与帧率（FPS）。为了在准确率与速度之间给出可比较的折衷指标，本文定义了复合评分Score=mIoU+ln(1+FPS+1/FLOPs)，用以在保留mIoU主导地位的同时，适度奖励推理速度的提升。

实验结果显示，单尺度下虽有较高的帧率（FPS=6.8）但平均mIoU较低（37.9%）；两尺度组合将平均mIoU提升到38.8%，FPS降至2.6；三尺度组合在mIoU上取得了略高的峰值（39.0%），但FPS进一步降低到1.9。考虑到实际部署对吞吐率的要求，我们采用包含速度项的复合指标进行权衡，按该指标计算后，两尺度组合[0.75, 1.25]的Score为39.7，略高于三尺度的39.6，因此被选为最终的尺度配置。

**表4　多视角一致性推理中尺度的调参实验**

**Table 4　Hyperparameter tuning experiments at the medium scale in Multi-view Consistency Reasoning**

| Scales | VOC | Context | COCO-Stf | ADE | City | Avg | FPS ↑ | FLOPs ↓ | Score ↑ |
|---|---|---|---|---|---|---|---|---|---|
| 1.0 | 64.4 | 36.9 | 26.7 | 20.1 | 41.1 | 37.9 | **6.8** | **4.1** | 40.0 |
| **0.75, 1.25** | **66.5** | 37.7 | 27.5 | 20.6 | 41.8 | 38.8 | 2.6 | 19.1 | **40.1** |
| 1.0, 0.75, 1.25 | 66.5 | **37.8** | **27.6** | **20.8** | **42.3** | **39.0** | 1.9 | 23.2 | 40.0 |

注：Score= mIoU+ln(1+FPS+1/FLOPs)，加粗字体为每行最优值。

表5对多视角一致性推理中的翻转策略与翻转后预测的融合方法进行了系统比较，首先在翻转类型上进行对比，包含不使用翻转、垂直翻转、不确定性像素翻转以及水平翻转的单独测试，结果表明水平翻转在平均mIoU指标上取得了最高值38.8%的Avg，而垂直翻转反而导致性能下降（Avg=37.7%），不确定性像素翻转与不使用翻转的效果接近。由于各策略的FPS相差不大，我们以平均mIoU作为主判据进行选择。随后针对水平翻转情形比较了三种融合策略：简单平均、基于熵的加权以及取最大值融合，实验显示简单平均融合取得了最高的平均mIoU（38.8%），略优于最大值融合（38.7%）和熵加权融

合(38.6%)。基于上述结果,本文最终确定使用水平翻转+平均融合作为默认配置,该组合在精度上最优且实现简单、稳定,适合实际部署。

2.3.3　文本引导重校准的参数调优

本文进行了文本引导重校准中对软分配的温度参数的对比实验,见表6所示,比较了　=1.0,0.5,2.0三种设定在若干数据集上的影响。结果显示三种温度下的总体表现非常接近:当　=1.0与　=0.5时平均mIoU均为39.1%,而　=2.0时平均mIoU略升至39.2%。在按数据集观察时,　=2.0对Cityscapes与COCO-Stuff等若干数据集有小幅提升,而其他数据集变化可忽略不计。这组实验表明文本引导重校准对温度参数具有较强的鲁棒性,其软分配机制在一定范围内能够自适应各类场景下的相似度分布,不会因温度的调整而引起不稳定的类别响应,从而增

**表5　多视角一致性推理中翻转策略的调参实验**

**Table 5　Hyperparameter tuning experiments of flip strategies in Multi-view Consistency Reasoning**

| 翻转策略 | 融合策略 | VOC | Context | COCO-Stf | ADE | City | Avg | FPS ↑ | FLOPs ↓ |
|---|---|---|---|---|---|---|---|---|---|
| 无翻转 | | 65.8 | 37.6 | 27.4 | 20.6 | 41.6 | 38.6 | **5.2** | **9.6** |
| 垂直翻转 | | 65.3 | 35.7 | 26.3 | 19.4 | 41.8 | 37.7 | 2.6 | 19.1 |
| 不确定性像素翻转 | | 65.8 | 37.6 | 27.5 | 20.6 | 41.8 | 38.7 | 2.6 | 19.1 |
| **水平翻转** | **平均融合** | **66.5** | **37.7** | **27.5** | **20.6** | 41.8 | **38.8** | 2.6 | 19.1 |
| | 熵加权融合 | 65.7 | 37.5 | 27.5 | 20.6 | 41.7 | 38.6 | 2.6 | 19.1 |
| | 最大值融合 | 66.0 | 37.6 | 27.5 | 20.6 | **41.9** | 38.7 | 2.6 | 19.1 |

注:加粗字体为每行最优值。

**表6　文本引导重校准中温度的调参实验**

**Table 6　Hyperparameter tuning experiment on temperature in Text-Guided Recalibration**

| 温度(　) | VOC | Context | COCO-Stf | ADE | City | Avg |
|---|---|---|---|---|---|---|
| 1.0 | 66.6 | 37.8 | 28.2 | 21.0 | 42.4 | 39.1 |
| 0.5 | 66.5 | **37.9** | 28.1 | 20.9 | 42.4 | 39.1 |
| **2.0** | **66.6** | 37.8 | **28.2** | **21.0** | **42.5** | **39.2** |

注:加粗字体为每行最优值。

强了在零样本设置下的可靠性。

鉴于　=2.0在平均指标上略占优势且有助于抑制过度尖锐的相似度归一化,本工作在最终配置中将　=2.0作为默认温度设定。

## 3　结　论

本文提出TG-CLIP,一种基于冻结视觉—语言模型的免训练开放词汇分割框架,旨在为具身智能与人形机器人等需要在真实环境中即时感知与决策的系统提供更稳健的像素级语义理解能力。针对CLIP在密集预测上的局部语义弱化及对尺度与视角的敏感性,本研究从推理端出发设计了两类轻量算子:文本引导重校准,在单次前向中将文本查询作为条件信号对patch级表征进行语义性重构与回投;多视角一致性推理,通过输入端的多尺度重采样与输出端的镜像融合放大跨尺度与跨视角的一致响应。所有改进均在前向路径完成,无需像素级微调或额外训练数据,在保持零样本泛化能力的同时,显著提升了像素级分割的判别质量与边界一致性。该设计具有良好的工程适配性,免训练推理算子可在受限计算条件下直接部署于具身智能与人形机器人系统。

实证上,TG-CLIP在八个开放词汇分割基准上均取得明显提升(见表1),综合性能优于多种先进方法。消融实验显示,多视角一致性推理与文本引导重校准各自能够带来稳定的性能增益,且在超参数方面表现出较好的鲁棒性,便于在不同数据域与

部署约束下进行工程化折衷。

尽管取得了可观进展，工作仍存在若干局限和值得改进的方面。首先，TG-CLIP依赖CLIP的patch分辨率，面对极小目标或需要精细边界的任务时仍有欠缺；其次，滑窗与多尺度策略在提升精度的同时带来了推理开销，影响在资源受限场景的实时部署；第三，本方法在语义高度相近类别的区分上仍依赖提示设计与类别映射策略，缺乏通过学习自适应修正的能力。以上问题主要源于免训练范式的适应性限制及基础表征的分辨率瓶颈。

基于上述分析，本文指出了若干可行的后续工作方向：一是研究轻量化的边界细化与局部放大算子，在保持免训练原则下进一步提升精细化性能；二是通过蒸馏或剪枝降低多视角一致性推理的计算开销，使多尺度策略更适用于实时或嵌入式场景；三是将文本引导重校准与自蒸馏或原型精炼结合，增强长尾类别判别能力。最后，考虑到实际应用需求，可进一步将TG-CLIP扩展至交互式分割与多任务场景，以评估其系统级应用价值。

综上，TG-CLIP在保持零样本开放性的同时，通过两类轻量化、前向可执行的改进，显著提升了像素级分割的局部语义一致性与尺度、视角鲁棒性。本文既提供了可直接部署的实践方案，也为后续在免训练范式下进一步缩小密集预测与基础VLM表征差距的研究指明了方向。

## 参考文献(References)

Bai S, Liu Y, Han Y, Zhang H, Tang Y, Zhou J and Lu J. 2025. Self-Calibrated CLIP for Training-Free Open-Vocabulary Segmentation. IEEE Transactions on Image Processing, 34: 8271-8284 [DOI: 10.1109/TIP.2025.3639996]

Bai X F, Lu L B and Wang W J. 2024. Saliency guided object complementary hiding for weakly supervised semantic segmentation. Journal of Image and Graphics, 29(4): 1041-1055 (白雪飞, 卢立彬, 王文剑. 2024. 显著性引导的目标互补隐藏弱监督语义分割. 中国图象图形学报, 29(04): 1041-1055) [DOI: 10.11834/jig.230156]

Bousselham W, Petersen F, Ferrari V and Kuehne H. 2024. Grounding Everything: Emerging Localization Properties in Vision-Language Transformers//Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024: 3828-3837 [DOI: 10.1109/CVPR52733.2024.00367]

Caesar H, Uijlings J and Ferrari V. 2018. COCO-Stuff: Thing and Stuff Classes in Context//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, UT, USA: IEEE/CVF, 2018: 1209 - 1218. [DOI: 10.1109/CVPR.2018.00132]

Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S and Schiele B. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE/CVF, 2016: 350.

Ding Z, Wang J and Tu Z. 2023. Open-Vocabulary Universal Image Segmentation with MaskCLIP//Proceedings of the 40th International Conference on Machine Learning. Honolulu: Proceedings of Machine Learning Research, Vol. 202: 8090 - 8102 [DOI: 10.5555/3618408.3618729].

Dong X, Zheng Y, Bao J, Zhang T, Chen D D, Yang H, Zeng M, Zhang W, Yuan L, Chen D, Wen F and Yu N. 2023. MaskCLIP: Masked self-distillation advances contrastive language-image pre-training//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE/CVF: 10995 - 11005 [DOI:10.1109/CVPR52729.2023.01059]

Everingham M, Van Gool L, Williams C K I, Winn J and Zisserman A. 2010. The PASCAL Visual Object Classes (VOC) Challenge. International Journal of Computer Vision, 88(2): 303 - 338. [DOI: 10.1007/s11263-009-0275-4]

Ghiasi G, Gu X, 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: Springer: 540 - 557 [DOI: 10.1007/978-3-031-20059-5_31].

Hajimiri S, Ben Ayed I and Dolz J. 2025. Pay Attention to Your Neighbours: Training-Free Open-Vocabulary Semantic Segmentation// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). 2025: 5061-5071.

Kang D and Cho M. 2025. In Defense of Lazy Visual Grounding for Open-Vocabulary Semantic Segmentation//Computer Vision — ECCV 2024. Cham: Springer Nature Switzerland, 2025: 143-164 [DOI: 10.1007/978-3-031-72940-9_9]

Kombol N. 2025. A Survey on Training-free Open-Vocabulary Semantic Segmentation[EB/OL].[2025-10-13]. https://arxiv.org/abs/2505.22209

Lan M, Chen C, Ke Y, Wang X, Feng L and Zhang W. 2025. ProxyCLIP: Proxy Attention Improves CLIP for Open-Vocabulary Segmentation//Computer Vision — ECCV 2024. Cham: Springer Nature Switzerland, 2025: 70-88. ISBN: 978-3-031-73113-6

Li B, Weinberger K Q, Belongie S, Koltun V and Ranftl R. 2022. Language-driven Semantic Segmentation (LSeg)//Proceedings of the 10th International Conference on Learning Representations. Online: ICLR / OpenReview [EB/OL].[2025-10-13]. https://arxiv.org/abs/2201.03546

Li Y, Wang H, Duan Y, Zhang J and Li X. 2025. A closer look at the

explainability of Contrastive language-image pre-training. Pattern Recognition, 162: 111409 [DOI: 10.1016/j.patcog.2025.111409]

Lin T Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P and Zitnick C L. 2014. Microsoft COCO: Common Objects in Context//Computer Vision — ECCV 2014 (Lecture Notes in Computer Science, vol. 8693. Cham, Switzerland: Springer, 2014: 740 - 755. [DOI: 10.1007/978-3-319-10602-1_48]

Liu Y, Bai S, Li G, Wang Y and Tang Y. 2024. Open-vocabulary segmentation with semantic-assisted calibration//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE, 3491 - 3500

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, HI, USA: PMLR, 956: 1 - 12

Luo J, Khandelwal S, Sigal L and Li B A. 2024. Emergent open-vocabulary semantic segmentation from off-the-shelf vision-language models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE/CVF, 4029 - 4040

Mottaghi R, Chen X, Liu X, Cho N G, Lee S W, Fidler S, Urtasun R and Yuille A. 2014. The role of context for object detection and semantic segmentation in the wild//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA: IEEE, 891 - 898 [DOI: 10.1109/CVPR.2014.119]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G and Sutskever I. 2021. Learning transferable visual models from natural language supervision//Proceedings of the 38th International Conference on Machine Learning. PMLR: 8748 - 8763

Shin H, Kim C, Hong S, Cho S, Arnab A, Seo P H and Kim S. 2024. Towards Open-Vocabulary Semantic Segmentation Without Semantic Labels[EB/OL]. [2025-10-14]. https://arxiv.org/abs/2409.19846

Stegmüller T, Lebailly T, Đukić N, Bozorgtabar B, Tuytelaars T and Thiran J-P. 2024. A Simple Framework for Open-Vocabulary Zero-Shot Segmentation (SimZSS)[EB/OL].[2025-10-15]. [https://arxiv.org/abs/2406.16085]https://arxiv.org/abs/2406.16085

Wang F, Mei J and Yuille A. 2025. SCLIP: Rethinking self-attention for dense vision-language inference//Proceedings of the European Conference on Computer Vision (ECCV). Cham: Springer Nature Switzerland: 315 - 332 [DOI: 10.1007/978-3-031-72664-4_18]

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

Xiang W K, Zhou Q, Cui J C, Mo Z Y, Wu X F, Ou W H, Wang J D and Liu W Y. 2024. Weakly supervised semantic segmentation based on deep learning. Journal of Image and Graphics, 29(5): 1146-1168 (项伟康, 周全, 崔景程, 莫智懿, 吴晓富, 欧卫华, 王井东, 刘文予. 2024. 基于深度学习的弱监督语义分割方法综述. 中国图象图形学报, 29(05):1146-1168) [DOI:10.11834/jig.230628]

Xu J, De Mello S, Liu S, Byeon W, Breuel T, Kautz J and Wang X. 2022. GroupViT: Semantic Segmentation Emerges from Text Supervision[EB/OL].[2025-10-16]. https://arxiv.org/pdf/2202.11094.pdf

Yuan H, Li X, Zhou C, Li Y, Chen K and Loy C C. 2024. Open-Vocabulary SAM: Segment and Recognize Twenty-thousand Classes Interactively//Proceedings of the European Conference on Computer Vision. Cham: Springer: 419 - 437. [DOI:10.1007/978-3-031-72775-7_24]

Yu Q, He J, Deng X, Shen X and Chen L C. 2023. Convolutions Die Hard: Open-Vocabulary Segmentation with Single Frozen Convolutional CLIP//Proceedings of the 37th Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates, Inc.: [DOI:10.5555/3666122.3667521]

Zhang H, Li F, Zou X, Liu S, Li C, Yang J and Zhang L. 2023. A Simple Framework for Open-Vocabulary Segmentation and Detection//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris: IEEE: 1020 - 1031. [DOI: 10.1109/ICCV51070.2023.00100]

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

Zhou Z, Lei Y, Zhang B, Liu L and Liu Y. 2023. ZegCLIP: Towards Adapting CLIP for Zero-Shot Semantic Segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE: 11175 - 11185 [DOI: 10.1109/CVPR52729.2023.01075].