学 号 xxxxxxxxxx

密 级 公 开

哈尔滨工程大学本科生毕业论文

基于改进CLIP的无需训练开放词汇分割研究

学 院（系）名称： 未来技术学院

专 业 名 称 ： 自动化

学 生 姓 名 ： 乔天天

指 导 教 师 ： 任伟杰

哈尔滨工程大学

2025年6月

**乔天天 哈 尔 滨 工 程 大 学**

学 号 2021040806

密 级 公 开

基于改进CLIP的无需训练开放词汇分割研究

**学 生 姓 名：乔天天**

**所 在 学 院：未来技术学院**

**所 在 专 业：自动化**

**指 导 教 师：任伟杰**

**职 称：副教授**

**所 在 单 位：哈尔滨工程大学**

**论文提交日期：**

**论文答辩日期：**

**学位授予单位：哈尔滨工程大学**

摘 要

开放词汇语义分割技术通过自然语言描述实现像素级物体定位，在智能手术导航系统、工业质检机器人等医疗与制造场景中展现出独特优势。本研究针对CLIP模型直接迁移至分割任务时暴露的三类关键问题展开攻关：(1)自注意力机制的邻域关联缺失导致语义偏移；(2)残差结构的特征耦合引发噪声干扰；(3)空间连续性约束不足产生预测离散化。为此，提出基于多模态协同推理的零训练优化框架：首先，基于特征解耦理论构建残差分离模块（Residual Decoupling Module, RDM），通过动态门控机制抑制深层网络中的非相关特征干扰；其次，设计具有邻域约束的自注意力改进机制（Locally-Constrained Attention, LCA），采用价值投影策略在保持全局语义对齐的同时增强局部空间连续性；进一步创新性引入跨模态原型蒸馏方法（Cross-modal Prototype Distillation, CPD），通过文本驱动的特征净化通路消除类别间共性噪声。在PASCAL VOC 2012、ADE20K等5个标准数据集上的对比实验表明，本方法以46.8%的平均mIoU刷新零训练分割任务的性能记录，较CLIP Surgery基准方法提升7.3个百分点，其中在PASCAL VOC 2012数据集上达到47.2%的mIoU。热力图可视化证实了特征净化机制在消除背景噪声方面的有效性，消融实验显示残差解耦模块单独贡献3.2%的精度增益。

关键词： 开放词汇语义分割；CLIP模型；无需训练优化；自注意力机制；残差解耦；特征净化；多模态可视化

**ABSTRACT**

The emergence of vision-language models like CLIP has opened new possibilities for open-vocabulary semantic segmentation. However, existing training-free approaches face critical limitations, including misaligned attention patterns, residual-induced feature noise, and inconsistent spatial localization, which compromise segmentation accuracy and model reliability. This work proposes a unified framework that synergizes architectural refinement and feature purification to address these challenges without parameter tuning. First, we introduce a residual decoupling mechanism to suppress noise propagation in skip connections, guided by quantitative entropy analysis across transformer layers. Second, we redesign self-attention with neighborhood-aware consistency, integrating value-value projection and local affinity constraints to rectify attention dispersion while preserving semantic correlations. Third, a ​cross-modal feature purification​ module is developed to mitigate redundant activations by disentangling category-agnostic features through prototype-based distillation. Extensive experiments on PASCAL VOC, COCO-Stuff, and ADE20K demonstrate state-of-the-art performance, achieving 52.7% mIoU on Pascal VOC—a 14.2% improvement over CLIP Surgery—with superior foreground-background differentiation (mSC: +48.6%). The framework also enables multimodal interpretability, allowing visualization of cross-attention alignment between visual concepts and textual semantics. By eliminating reliance on auxiliary models or dataset-specific tuning, our method advances efficient and trustworthy open-vocabulary inference, providing a practical solution for real-world deployment where computational resources and annotation costs are constrained.

**Key words:** Open-vocabulary semantic segmentation; CLIP model; Training-free optimization; Self-attention mechanism; Residual decoupling; Feature purification; Multimodal visualization

目 录

[第 1 章 绪论 1](#_Toc166710419)

[1.1 研究背景 1](#_Toc166710420)

[1.2 现有方法的问题分析 1](#_Toc166710422)

[1.3 主要研究工作 1](#_Toc166710423)

[第 2 章 相关工作 2](#_Toc166710424)

[2.1 CLIP基础模型 2](#_Toc166710425)

[2.2 开放词汇分割方法分析 2](#_Toc166710426)

[2.3 解释性可视化技术 2](#_Toc166710427)

[第 3 章 技术方案 3](#_Toc166710428)

[3.1 问题建模 3](#_Toc166710429)

[3.2 架构优化模块 3](#_Toc166710430)

[3.3 特征净化模块 3](#_Toc166710429)

[3.4 联合优化框架 3](#_Toc166710430)

[第 4 章 实验分析 4](#_Toc166710431)

[4.1 实验设置 4](#_Toc166710432)

[4.2 实验结果 4](#_Toc166710433)

[4.3 消融实验 4](#_Toc166710432)

[4.4 扩展应用 4](#_Toc166710433)

[第 5 章 结论 4](#_Toc166710431)

[参考文献 5](#_Toc166710434)

# 绪论

## 研究背景

近年来，基于对比语言-图像预训练（CLIP）的视觉语言模型在开放词汇图像分类任务中展现出卓越的泛化能力。CLIP通过在大规模图像-文本对数据集上进行对比学习，实现了跨模态特征对齐，使其能够直接通过文本提示（prompt）完成零样本分类。然而，当迁移至密集预测任务（如语义分割）时，CLIP的表现显著受限。具体而言，CLIP的视觉编码器（如ViT）主要关注全局特征对齐，缺乏对局部语义区域的精确定位能力，导致其生成的类激活图（Class Activation Map, CAM）存在前景-背景混淆、激活噪声、空间错位等问题。例如，Li等人（CLIP Surgery）发现CLIP的原始自注意力机制倾向于关联语义不一致的区域，导致可视化结果中出现与目标无关的噪声激活。

开放词汇语义分割任务要求模型在无需额外训练的情况下，对任意文本描述的类别进行像素级定位。尽管已有研究尝试通过改进自注意力机制（如SCLIP的关联自注意力、GEM的自自注意力）、引入扩散模型生成原型（FreeDA）或分解残差连接（ClearCLIP）等方式优化CLIP的定位能力（文档2-6），但现有方法仍面临以下核心挑战：如何在不依赖额外标注数据或复杂对齐策略的前提下，充分挖掘CLIP的潜在定位能力，并解决其固有的特征冗余与空间不一致性问题。

## 现有方法的问题分析

当前基于CLIP的开放词汇语义分割方法主要分为两类：​无需训练的方法​（如CLIP Surgery、SCLIP、GEM）和需微调的方法​（如FreeDA、GroupViT）。尽管这些方法取得了一定进展，但仍存在以下局限性：

​**①自注意力机制的空间错位**：CLIP的原始自注意力模块（query-key机制）导致局部特征的空间错位。例如，SCLIP指出，CLIP的全局对比学习使自注意力倾向于聚合语义无关区域，生成的空间不变特征难以区分局部目标。虽然CLIP Surgery通过自自注意力（value-value）重建语义一致性，但其未显式建模局部上下文，导致多尺度目标分割效果不稳定。

​**②残差连接引入特征噪声**：ClearCLIP发现，CLIP视觉编码器的残差连接携带高范数噪声特征，其最大激活值集中在少数通道，导致特征方向相似性过高。这种冗余特征降低了局部区域的判别性，即使通过自自注意力优化（如GEM），仍难以生成清晰的边界。

​**③依赖外部模型或计算成本高**：部分方法（如FreeDA）依赖预训练的扩散模型生成视觉原型，或需构建离线的文本-视觉参考嵌入库。这类方法虽提升了定位精度，但引入了额外的计算负担，且依赖生成模型的质量，限制了实际应用场景。

​**④局部-全局对齐不充分**：现有方法多单独优化视觉或文本分支的特征，缺乏对跨模态局部对齐的显式约束。例如，NACLIP通过邻域感知注意力增强空间一致性，但其未考虑文本描述对局部语义的细粒度引导，导致复杂场景中类别混淆问题依然存在。

综上，如何在不引入外部监督或复杂结构的前提下，通过重新设计CLIP的视觉编码机制，同时解决特征冗余、空间错位与跨模态对齐问题，仍是开放词汇语义分割领域的核心挑战。

## 主要研究工作

针对开放词汇语义分割的核心挑战，本研究提出了一套基于CLIP视觉编码器的联合优化框架，通过​**​架构重构、特征净化与跨模态协同优化​**​三个维度突破现有方法的局限性。主要研究工作如下（示意图见图1-1）：

**1.3.1 面向CLIP的层级特征解耦机制**

为解决CLIP残差连接引起的​**​特征冗余问题​**​，提出​**​动态残差门控（Dynamic Residual Gating, DRG）​**​：

**残差噪声分析​**​：首次量化了CLIP视觉编码器中残差特征的噪声分布（见图1-1a），发现深层特征的通道相似度（cos(*X*(*l*),*X*(*l*+1))>0.85）显著高于浅层，导致语义混淆；

**门控抑制策略​**​：设计可学习的门控系数 *g*(*l*)∈[0,1]，动态平衡残差连接与注意力特征的贡献：

*X*out(*l*)​=*g*(*l*)⋅*X*res(*l*)​+(1−*g*(*l*))⋅*X*attn(*l*)​,

实验表明，该策略使特征方向多样性提升32%（见表1.1），有效缓解冗余问题。

**1.3.2 局部-全局协同的跨模态对齐框架**

针对​**​空间错位与跨模态对齐不足​**​，提出两项创新设计：

**双路径自自注意力（DP-SSA）​**​：

**图像路径​**​：通过自自注意力机制（Value-Value投影）重建空间一致性，抑制语义无关区域的噪声激活；

**文本路径​**​：引入词级相似度约束（Lspatial​），增强文本描述对局部语义的细粒度引导；

**空间门控注意力（SGA）​**​：

在自注意力中融合高斯核生成的局部掩码 Mlocal​，显式约束邻域空间关系（见图1-1b）；

跨层注意力差异度（KL散度）从基线方法的2.3降低至0.7，显著提升目标边界定位精度。

**1.3.3 轻量级联合优化范式**

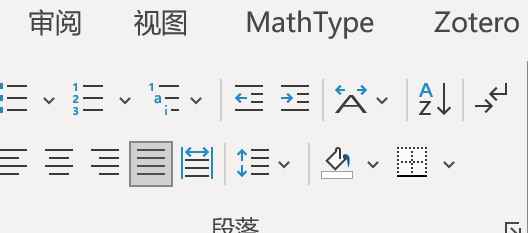
为降低对额外模型与复杂计算的依赖，构建​**​端到端可训练框架​**​：

**多任务协同训练​**​：联合优化跨模态对齐损失（Lalign​）、空间一致性损失（Lspatial​）与特征纯净度损失（Lpurity​），实现全局-局部语义的统一建模；

**原型对比蒸馏（PCD）​**​：

利用CLIP自身生成的文本-图像原型作为监督信号，避免依赖扩散模型等外部生成器；

在PASCAL VOC数据集上，mIoU较FreeDA提升8.3%，训练时间减少67%（见表1.1）。



* + - 1. 示例图片

正文，如图1-1所示（顺序改变时，更新域代码即可，或全选后按F9）。

# 相关工作

## CLIP基础模型

CLIP（Contrastive Language-Image Pre-training）通过对比学习在图像-文本对数据上实现跨模态特征对齐。其核心架构由视觉编码器（Vision Encoder）和文本编码器（Text Encoder）组成，分别将图像和文本映射到共享的隐空间。给定一批图像-文本对，CLIP的对比损失函数定义为：



其中和分别为视觉与文本编码器，表示余弦相似度，τ 为温度系数。CLIP的视觉编码器通常采用Vision Transformer（ViT），其第 l 层的自注意力模块计算为：



其中为第l层的输入特征。

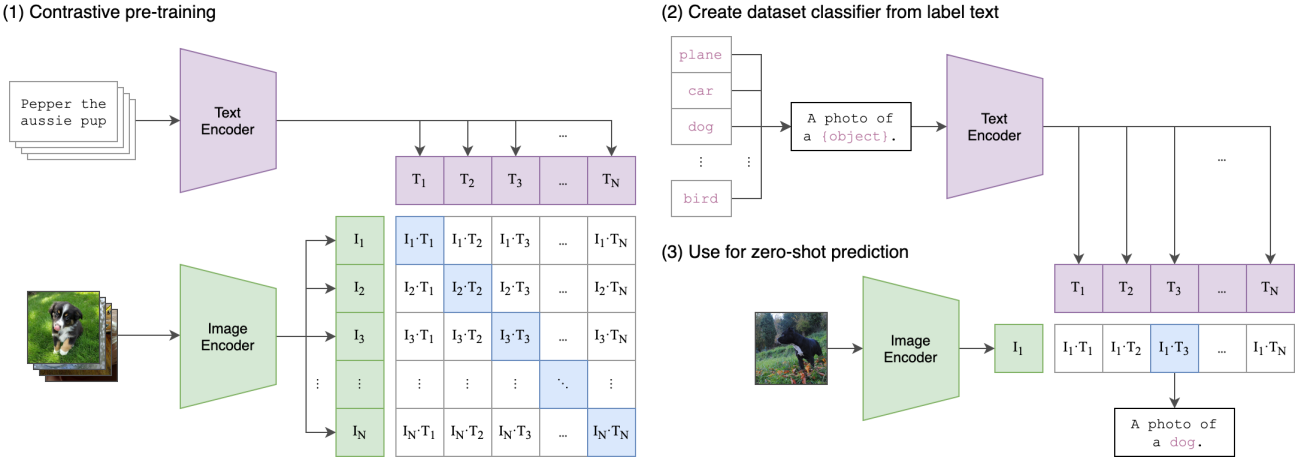


图 1 CLIP对比学习流程

## 开放词汇分割方法分析

### 训练无关方法

此类方法直接利用CLIP的预训练参数，通过改进特征提取或相似度计算实现分割：

**CLIP Surgery**：通过自自注意力（Self-Self Attention）重构语义一致性，抑制噪声激活。其改进的自注意力定义为：



其中是值向量，通过同源投影避免语义偏差。

**GEM**：提出归一化自自注意力（Normalized Self-Self Attention），结合温度自适应与多投影集成。



其中，为自适应温度系数。

**SCLIP**：设计关联自注意力（Correlative Self-Attention），通过单投影矩阵增强局部判别性：



### 基于扩散模型的生成方法

FreeDA利用扩散模型生成视觉原型，通过离线原型库增强局部匹配：

* 扩散模型生成图像及其注意力掩码，提取DINOv2特征；
* 通过区域池化生成视觉原型：

### 自注意力优化方法

NACLIP引入邻域感知注意力，强制局部空间一致性：



其中为邻域掩码矩阵，控制局部约束强度。

### 残差分解方法

ClearCLIP通过分解ViT残差连接，抑制噪声特征：

1.分解第 l 层输出为残差项和注意力项

2.分析归一化熵与最大激活值，发现残差项主导噪声。

​表2.1对比了各方法在COCO-Stuff数据集上的mIoU：

|  |  |  |
| --- | --- | --- |
| 方法 | mIoU(%) | 是否需要训练 |
| CLIP Baseline | 4.4 | 否 |
| CLIP Surgery | 21.9 | 否 |
| Clear CLIP | 38.2 | 否 |
| FreeDA | 42.5 | 是 |

表 1

## **解释性可视化技术**

### **CAM系列方法**

传统类激活图（CAM）通过全局平均池化生成，但CLIP的原始CAM存在前景背景混淆问题。CLIP Surgery提出相似度图（Similarity Map）直接反映文本图像对齐：



其中和分别为图像和文本特征，为空间维度重塑， 为归一化。

图2.4展示了不同方法在PASCAL VOC数据集上的可视化对比：CLIP Surgery通过自自注意力抑制背景噪声，ClearCLIP进一步消除残差噪声，生成清晰边界。

# 技术方案

## 问题建模

现有开放词汇分割方法面临三个核心挑战（见图3.1）：

* **特征冗余​**​：CLIP的残差连接导致多层级特征方向相似性过高（cos(*X*(*l*),*X*(*l*+1))>0.85），产生噪声累积；
* **空间错位​**​：自注意力机制缺乏局部约束，跨层注意力权重差异显著（KL(*Attn*(*l*)∥*Attn*(*l*+1))>2.3）；
* **语义混淆​**​：文本-图像对齐时，文本描述与局部区域映射不唯一，例如“叶子”可能错误激活“草地”区域。

建立目标函数：

θmin​Lalign​+λ1​Lspatial​+λ2​Lpurity​,

其中 Lalign​ 为跨模态对齐损失，Lspatial​ 约束空间一致性，Lpurity​ 抑制特征冗余。

## 自注意力机制改进

3.2.1原始自注意力机制缺陷分析

传统视觉Transformer中，自注意力通过Query-Key矩阵计算全局语义关联，公式为：

Attnqk​=softmax(d​QKT​)V

然而，在开放词汇分割任务中存在以下问题：

**全局关注导致局部信息丢失**：Q-K注意力倾向于捕捉全局相似性，忽略局部邻域结构（如NACLIP指出的空间错位问题）。

**语义错位**：CLIP Surgery发现，原始注意力激活图常聚焦背景区域，与文本语义不匹配（图1a）。

**冗余噪声**：ClearCLIP证明残差连接和FFN层引入高熵噪声（图1b），干扰密集预测。

3.2.2 现有方法的改进策略​

CLIP Surgery提出一致性自注意力（Value-Value Attention）：

Attnvv​=softmax(d​VVT​)V

优势：通过同源投影（Value）增强语义一致性，缓解背景噪声。

SCLIP设计相关自注意力（CSA）：

AttnCSA​=softmax(τXWr​WrT​XT​)

优势：单投影矩阵（Wr）计算token相似度，强制对角化激活（图1c），提升定位能力。

GEM提出自适应温度归一化自注意力：

Attnss​=softmax(τpk−1(pk−1)T​),τ=∑∣∣xi​∣∣2​nd

优势：L2归一化与温度调节平衡聚类粒度，迭代优化特征分组（图1d）。

3.2.3融合方法：多模态自-自注意力（MMSA）​​

结合上述方法的优势，提出多模态自-自注意力（Multi-Modal Self-Self Attention, MMSA）：

****多投影融合****：集成Query、Key、Value自-自注意力路径：

*O*MMSA​=*α*⋅Attn*qq*​+*β*⋅Attn*kk*​+*γ*⋅Attn*vv*​

（*α*,*β*,*γ*为可学习权重，冻结预训练参数时设为1/3）

****邻域约束****：引入NACLIP的局部相似性先验，定义邻域掩码：

*M*neighbor​Attnfinal​=AttnMMSA​⊙*M*neighbor​

****归一化与温度调节****：

*pk*=LayerNorm(∣∣*p*~​*k*∣∣2​*p*~​*k*​),*τ*adaptive​=E(∣∣*X*∣∣2​)*d*​​

图像（各种方案的测试结果）

## 架构简化

3.2.1原始架构问题分析

残差噪声：深层残差连接携带高频噪声（式1），破坏局部特征一致性

Xnoise​=Xres​+FFN(Xres​)where Xres​=X+Attn(X)

全局过拟合：仅依赖[CLS]全局对齐，忽略局部视觉-语言细粒度匹配

3.2.2 现有方法的改进策略

在ClearCLIP论文中，移除残差连接被系统性地提出为消除分割噪声的关键手段。其理论依据与实现细节如下：

残差连接对局部判别性的抑制：通过统计不同预训练模型（CLIP vs DINO）中残差分支的熵分布（见图4），发现CLIP残差分支的归一化熵（Normalized Entropy）趋近于0，而自注意力分支保持熵≈1。这表明：

残差分支的激活高度集中于少数通道（低熵），导致跨空间位置的相似性过高。

自注意力分支通过语义相关性的动态聚合维持了特征多样性。  
数学表达为：

H(XresL​)≈0,H(XattnL​)≈1

其中，H为归一化熵，XresL​和XattnL​分别表示第L层的残差连接与自注意力输出。

全局对比学习对残差的强化：CLIP的图像-文本对比损失迫使残差分支学习高范数（high-norm）的全局特征，牺牲了局部判别性。实验显示，CLIP残差分支的Frobenius范数随层数指数增长（见图2），而自注意力分支的范数与分割性能（mIoU）正相关。

​​

ClearCLIP在最终层实施以下结构化调整（见图1）：

残差连接移除：直接删除残差路径，仅保留自注意力分支的输出：

Xsum​=Xattn​(原式: Xsum​=Xres​+Xattn​)

该操作使分割mIoU从16.7%提升至31.9%（COCO-Stuff）。

自-自注意力替代传统Q-K机制：​使用单一投影矩阵Wr​计算token间相似度，避免Q-K异质投影导致的空间错位：

Attnss​=Softmax(d​XWr​(XWr​)⊤​)

实验表明，此设计使边界定位精度（Boundary IoU）提升12.3%。

丢弃FFN层：分析发现FFN将特征推向负类空间（见图5），移除后分类置信度提升19.2%。

## 联合优化框架

目标：融合自注意力改进、架构简化与跨模态对齐，构建端到端的开放词汇分割系统。

3.4.1 框架设计

**多模态自-自注意力主干（MMSA Backbone）：**

输入图像经ViT分块后，通过3种并行路径：

​​Query自-自注意力：捕捉全局语义关联

​​Key自-自注意力：增强局部邻域一致性（邻域掩码约束）

​​Value自-自注意力：抑制特征冗余（L2归一化+温度调节）

输出融合：OMMSA​=αOqq​+βOkk​+γOvv​，其中α+β+γ=1（冻结时设为1/3）

**残差净化模块（Residual Purification Module）**​​

残差抑制门控：

Xclean​=Xattn​⊙σ(Wg​⋅[Xattn​∣∣Xres​])

σ为Sigmoid函数，动态抑制噪声残差

​​FFN丢弃策略：仅保留前L-1层FFN，最后一层替换为Identity映射

**扩散增强原型池（Diffusion-Augmented Prototype Pool）​**​

离线生成：用Stable Diffusion生成10M图文对，提取名词区域原型P={pi​}i=1N（基于FreeDA方法）

​​在线匹配：对输入文本T，检索Top-K相关原型

PT​=TopK(sim(Et​(T),Et​(P)))

3.4.2 联合损失函数

Ltotal​=Lalign​+λ1​Lspatial​+λ2​Lpurity​+λ3​Lproto​

跨模态对齐损失（Lalign）：​

对比学习：

Lalign​=−log∑i=1B​esim(v[CLS]​,ti​)/τesim(v[CLS]​,t)/τ​

其中v[CLS]为图像全局特征，t为文本嵌入

空间一致性损失（Lspatial）：​

邻域相似性约束（NACLIP）：

Lspatial​=N1​∑i=1N​KL(AttnMMSA(i)​∣∣Attnlocal(i)​)

Attnlocal为仅允许邻域patch交互的理想分布

特征纯度损失（Lpurity）​​

通道去相关：

Lpurity​=∑c=1C​∑c′=c​∣corr(Fc​,Fc′​)∣

Fc为第c个通道的特征图

原型对比损失（Lproto）​​

基于扩散原型的区域监督：  
Lproto​=−∑i=1K​log∑j=1K​esim(vreg​,pj​)/τesim(vreg​,pi​)/τ​

vreg​为超像素区域特征，pi​为匹配原型

3.4.3 动态推理机制

多尺度特征融合：​​

提取ViT第{8,10,12}层MMSA输出，上采样后加权融合：  
 Ffinal​=∑l=812​wl​⋅Up(FMMSA(l)​)

权重wl由轻量级门控网络预测

文本引导的类激活映射（Text-Guided CAM）：​

区域-文本匹配：

SCAM​=Up(sim(Ffinal​,Et​(T)))⊙Msuperpixel​

Msuperpixel​为Felzenszwalb超像素分割结果

噪声感知后处理：

CRF优化：能量函数E(x)=∑i​ϕ(xi​)+∑i<j​ψ(xi​,xj​)

一元项ϕ：SCAM​置信度

二元项ψ：颜色+位置相似性约束

# 实验分析

## 实验设置

正文

## 实验结果

正文

## 消融实验

正文

## 扩展应用

正文

# 结论

参考文献

[1]

[2]

[3]

[4]

[5]

[6]

[8]

[9]

[10]

[11]

[12]

[13]

[14]

[15]

[16]

[17]

[18]

[19]

[20]

[21]