TagCLIP：一个增强开放词汇的本地到全球框架无需训练的 CLIP 多标签分类

Yuqi Lin1,3, Minghao Chen2*, Kaipeng Zhang3*, Hengjia Li1 , Mingming Li1 ，
Zheng Yang4 , Dongqin Lv6 , Binbin Lin5 , Haifeng Liu1 , Deng Cai1,4
1浙江大学计算机学院CAD&CG国家重点实验室
2杭州电子科技大学3上海人工智能实验室4FABU Inc.
5浙江大学软件学院6南通港口集团
{linyq5566, minghaochen01}@gmail.com, kp.zhang@foxmail.com

抽象的

对比语言-图像预训练 (CLIP) 在开放词汇分类中展现了令人印象深刻的能力。图像编码器中的类别标记经过训练，能够捕捉全局特征，从而区分受对比损失监督的不同文本描述，这使得它对于单标签分类非常有效。然而，它的表现不佳

在多标签数据集上的表现，因为全局特征往往由最突出的类别主导，并且

softmax 运算的对比性质加剧了这种情况。在
在本研究中，我们观察到多标签分类结果严重依赖于判别性局部特征，而这些特征被 CLIP 忽略了。因此，我们分析了

CLIP 中的逐块空间信息，并提出了一个从局部到全局的框架来获取图像标签。它包含三个
步骤：（1）块级分类以获得粗略分数；
（2）双掩蔽注意力细化（DMAR）模块，用于重新细化粗略分数；（3）类别重新识别（CWR）
模块来从全局角度修正预测。这
该框架完全基于冻结的 CLIP，并且显著
无需针对特定数据集进行训练，即可在各种基准上提高其多标签分类性能。此外，
全面评估生成标签的质量和实用性，我们将其应用扩展到下游任务，

即弱监督语义分割（WSSS）
生成的标签作为图像级伪标签。实验
证明这种先分类后分割的范式显著优于其他无注释分割方法，并验证了生成的标签的有效性。我们的

代码可在 https://github.com/linyq2117/TagCLIP 上找到。

介绍

对比语言-图像预训练（CLIP）（Radford
等人（2021）最近提出了一个强大的视觉语言模型。它是在一个大规模数据集上进行预训练的
图像-文本对，并表现出令人印象深刻的性能
在图像文本匹配任务中（Zhou et al. 2022；Crowson
et al. 2022; Gu et al. 2021). 通过将这种匹配能力迁移到分类任务，我们可以识别任意文本标签并实现开放词汇分类。

然而，大多数现有的开放词汇著作都集中于

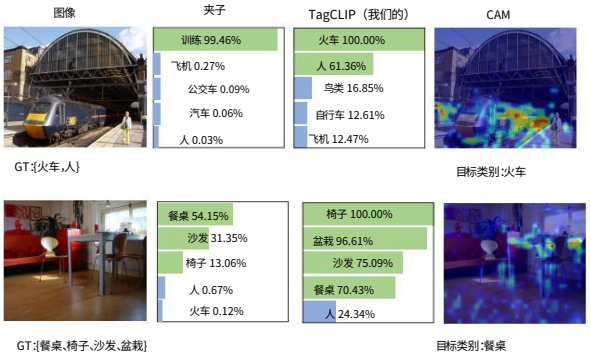

图 1：多标签分类结果的可视化
以及一些目标类别的 CAM。中间两列
证明原始 CLIP (Radford et al. 2021) 通常无法识别不显眼的类别，而我们的
TagCLIP 可以很好地识别它们。最后一列介绍了
一些特定类别的 CAM，并表明分类主要取决于一些判别性局部特征。

所有结果均基于 ViT-B/16，我们利用 Grad-CAM（Selvaraju 等人，2017）来获取 CLIP 的 CAM。

单标签分类任务，而多标签分类旨在识别所有相关类别或

概念在图像中，是一个更实际、更具有挑战性
任务。在图1中，我们发现在多标签分类数据集上的表现并不令人满意。具体来说，
由类别标记预测的分类逻辑倾向于
最突出的类别占主导地位，而一些不显眼的物体，例如尺寸较小的物体，通常被低估。这主要有两个原因：（1）CLIP 是经过训练的

使用对比损失来对齐图像-文本对，其目的是
将图像与其对应的文本描述进行匹配
并将其与其他模型区分开来。此损失函数引入的softmax操作在不同模型之间创造了竞争
类别，这对多标签设置是有害的。（2）
CLIP 经过训练，可以通过独特的
使用类标记进行全局嵌入，无需明确
捕捉特定区域的局部特征。然而，
多标签设置中，判别性局部特征更加
很有帮助。这种对本地特征的偏好可以在
类别激活图（CAM）（Zhou et al. 2016）

*通讯作者

在图 1 中,目标的高度响应区域
类别主要对应于特定的局部线索。因此,它
有必要探索保存的空间信息
CLIP-ViT 利用判别性局部线索。

一般来说,模型最终输出的特征图是
通常用于定位任务,例如物体检测（Ren et al. 2015)或分割（Chen et al. 2017）。
然而,我们观察到 CLIP-ViT 的定位质量对于最后一个特征图并不有效（见图 3）。
我们深入研究了背后的原因,发现最后几层的注意力操作对于密集的 token 来说是不合理的,
导致最终输出缺乏空间信息
特征图。或者,通过转发
倒数第二层,没有自注意力操作
最后一层（简称为倒数第二层）,
空间信息得到有效保留。这使我们能够
从 CLIP 中提取局部特征,增强其
用于捕捉细粒度的细节。

基于上述观察,我们进一步提出
名为 TagCLIP 的新框架可以增强多标签
无需训练即可达到原始 CLIP 的分类能力。该框架遵循从局部到全局的范式,并且
包含三个步骤。首先,我们忽略 CLIP-ViT 最后一层的注意力操作,并执行 patch-level
基于倒数第二层进行分类,得到每个类对应的分类得分图。其次,

为了改进初始分数并减轻潜在的噪音,我们引入了一种基于双掩蔽注意力改进策略
关于 ViT 固有的多头自注意力 (MHSA)。
最后,我们提出了一个类别重新识别模块来
进一步改进全球的主要预测
视图。这种双重检查方法可以过滤掉一些错误
检测类别并提高遗漏案例的分数。
整个框架显著提高了 CLIP 的多标签分类性能。它完全基于冻结的

CLIP 并支持开放词汇多标签分类
无需针对特定数据集进行训练。

为了进一步验证生成的标签的质量和实用性,我们将 TagCLIP 与下游任务集成,
它作为一个通用的注释器,提供
高质量的伪标签。它可以使许多下游
任务,例如自我训练（Zoph 等人,2020 年;Wang 等人）。
2022;Xie et al. 2020）,以及弱监督学习（Lin
等 2023;Xie 等 2022;Xu 等 2022b）。在本文中,我们通过整合

生成的标签与弱监督语义分割（WSSS)相结合。开放词汇表

多标签分类和 WSSS 实现无注释
分割。与之前的研究（Zhou、Loy 和 Dai
2022;Van Gansbeke 等人 2021)遵循自下而上的范式,我们惊讶地发现这种新颖的“先分类后分割”范式可以显着提高性能,

这表明了图像级监督的重要性
进行分割任务。

主要贡献可概括如下：

- 我们探索 CLIP 中的空间信息
  级别,并发现最后
  层打破了空间信息。在此基础上,我们提出了一个从局部到全局的框架TagCLIP来增强
  原始的多标签分类性能
  CLIP 无需任何额外培训。
- 实验结果证明了我们的有效性
  TagCLIP。它释放了原始 CLIP 的潜力
  并能生成高质量的图像标签。我们的方法
  与原始 CLIP 和其他不同基准测试中的作品相比,取得了显著的性能提升。
- 我们将建议的 TagCLIP 与下游
  WSSS 任务并找到这种先分类后分段的范式
  比其他方法取得了显着的进步。

相关作品

对比语言-图像预训练

对比语言-图像预训练（CLIP）（Radford
等人 2021)将视觉概念与文本描述联系起来,并为许多计算机视觉任务提供了

语言能力。它由图像和文本编码器组成,
并进行联合训练,使两种模式保持一致
4亿个图文对。图文匹配能力可以迁移到下游的零样本任务。

然而,预训练任务是图像级的,并且只有类
训练token是为了捕捉全局特征。对于多标签分类任务,区域级特征更受青睐。

一些作品（Raghu 等人 2021;Ghiasi 等人 2022)探索
深层 ViT 层中的空间信息,但
结果并不令人满意。本文通过
忽略上次的注意力操作,以及获得的杠杆
局部特征有利于多标签分类。

开放词汇多标签分类

多标签分类旨在预测一组标签
图像。传统上,多标签分类任务是
转化为一组二分类任务,
通过优化二元交叉熵损失函数来解决。提出的方法可以分为三类

主要方向:1)改进损失函数（Ridnik et al.
2021;Wu et al. 2020）。2)标签相关性建模（Chen
et al. 2019b,a; Ye et al. 2020). 3) 定位感兴趣区域 (Wang et al. 2017; You et al. 2020). 为了处理未见标签,开发了多标签零样本学习 (ML-ZSL),将知识从可见类迁移到未见类

类。这项任务的关键是图像的对齐
及其相关的标签嵌入以及
可见和不可见的标签嵌入。现有的研究成果实现了
从寻找主要方向（Ben-Cohen et al. 2021)或采用注意模块（Narayan

等人。 2021 年； Huynh 和 Elhamifar 2020）。

与 ML-ZSL 不同,视觉相关的语言数据
图像标题可以作为辅助监督
开放词汇设置。开放词汇多标签
图像识别可以通过任意文本名称或描述对多标签图像进行分类。根据复杂程度,现有方法可分为两类。1)第一类

小组需要对已见类进行额外的训练过程
或特定的精选数据。这些方法需要微调
在目标数据集上（He 等人 2023;Sun、Hu 和 Saenko
2022)或使用海量数据从头开始训练（Guo et al.
2023）,两者都有复杂的训练过程。2）

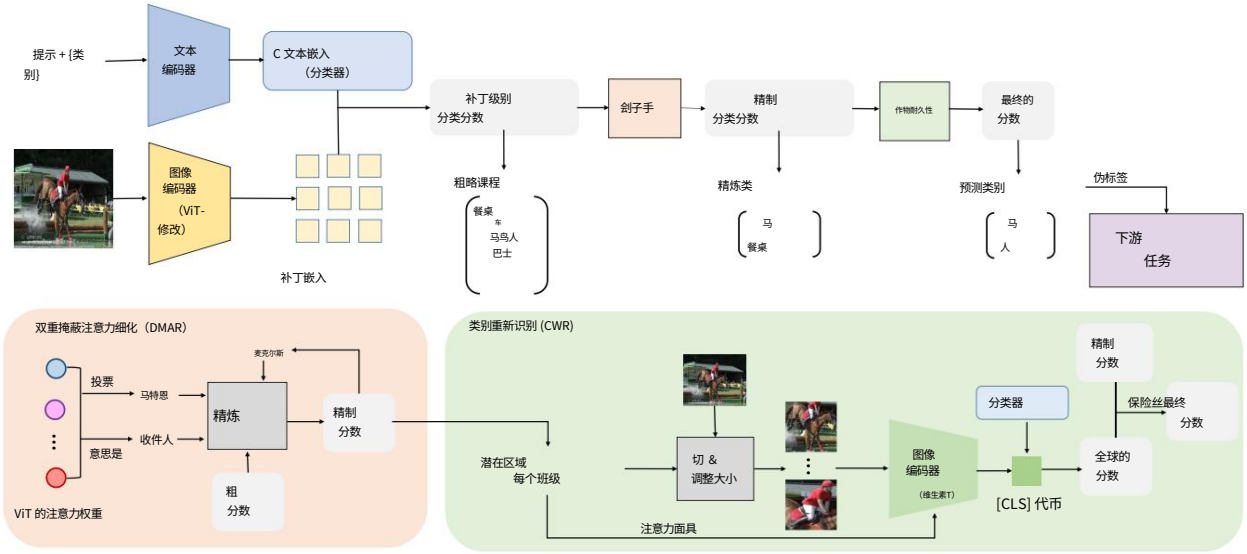

图 2:我们提出的框架概览。该框架包含三个步骤:块级分类、双掩蔽注意力细化 (DMAR) 和类别重识别 (CWR)。C 表示类别总数。“ViT-modified”表示忽略最后的自注意力操作以保留空间信息。我们以 0.5 为阈值对预测概率得分进行阈值化,以获得预测类别。预测的图像标签可作为下游任务（例如 WSSS)的伪标签。

第二组仅基于预先训练的模型,没有进一步的训练或额外的信息（Li et al. 2023）。

我们的工作属于第二类,难度更大,但使用起来更方便。与 CLIP-Surgery (Li et al. 2023) 类似,我们从模型可解释性的角度提升了 CLIP 的分类能力。不同之处在于,我们利用了从局部到全局的框架,而 CLIP-Surgery 仅依赖于全局嵌入。

无注释语义分割:在无注释分割环境中,训练过程中不提供任何

注释,这对应于无监督语义分割 (USS)。主要的 USS 方法利用自监督学习来学习像素级表示 (Ji, Henriques, and Vedaldi 2019; Cho et al. 2021; Ziegler and Asano 2022; Ke et al. 2022; Hwang et al. 2019; Van Gansbeke et al. 2021),然后可以使用学习到的表示通过 K 均值或线性分类器对图像片段进行聚类。这些自下而上的方法难以区分外观相似的不同类别,也难以识别外观各异的类别。

另一个类似的设置是开放词汇分割。

其目标是用文本描述的任意类别（而非固定的标记词汇)来分割图像。它通常通过对弱监督信号（例如图文对)进行训练来解决闭集限制（Xu et al. 2022a; Luo et al. 2023）。然而,近期的研究（Zhou, Loy, and Dai 2022; Shin, Xie, and Albanie 2022b,a)仅基于预训练的 CLIP,无需额外注释。由于缺乏高级语义指导,这些方法的性能提升仍然有限。我们将所有上述利用图文对或预训练模型的方法称为基于 CLIP 的方法。

方法

在本节中,我们介绍基于 CLIP 的多标签分类框架 TagCLIP,如图 2 所示。我们首先回顾 CLIP-ViT 的架构并研究在块中保存的空间信息。

然后,我们介绍了我们提出的无需标注和微调的局部到全局多标签分类框架。最后,我们展示了生成的图像标签在下游 WSSS 任务中的应用。

CLIP 分析CLIP（Radford

等人,2021)由一个图像编码器和一个文本编码器组成,并经过联合训练,用于将两种模态与大规模图像-文本对对齐。对于采用 Transformer 架构的图像编码器,预训练了一个 [cls] 标记来捕捉全局特征。给定具有 L 层的 ViT,最后一层 Transformer 的前向传播表达式如下:

$$X \quad L = XL-1 + a \qquad \text{左}, \tag{1}$$

$$= XL-1 + A \qquad L\,XL-1WL \qquad \text{在}, \tag{2}$$

$$\text{一个}^{\text{左}} \qquad \frac{(XL-1WL \quad \text{问})(XL-1WL)^{T}}{\sqrt{d}} \text{softmax}(\quad +\text{毫升}), \tag{3}$$

$XL = X \quad L + \text{MLP}(X \quad L)$，(4)其中XL－1表示 L-1 层的输出

token,MLP 表示 Transformer 模块中的自注意力和 MLP 模块。AL 对 L 层的注意力权重进行编码。σ表示 softmax 正则化,d 是XL－1的维度。ML是AL的注意力 mask。WQ 、WK、 WV是线性投影权重,用于生成

拉

,

MHSA 中的查询、键、值。XL由[cls] 令牌组成
和剩余的标记（表示为密集标记）：

$$XL = [x_{cls}^{左}, xL]. \quad (5)$$

正如引言中提到的，对比损失和
原始 CLIP 中的全局嵌入会损害多标签分类。或者，区域级特征

更适合识别图像中的多个类别。由于在对比预训练中仅使用了 [cls] 标记，原始
CLIP 的定位能力

弱（Zhong et al. 2022）。主要表现
将预训练的 CLIP 模型应用于
定位任务（例如，分割任务的 mIoU 仅为 16.2%
通过利用表 1 中的最终输出特征图）。

我们假设空间信息保留
在前一层的 CLIP 特征图中，但缺乏
最后一层的原因如下：(1) 查询和
最后注意层中的 key 仅参与优化 [cls] token 以执行加权和运算

并在预训练过程中将信息全球化。这是针对[cls] token 的特殊设计，但毫无意义且冗余
剩余的密集标记。(2)[cls] 标记在整个视觉转换器中起着相对较小的作用，并且
直到最后一层才用于全球化（Ghiasi 等人
2022）。因此，它几乎不会影响前几层的局部特征。为了验证这一点，我们使用了 12 层的 ViT-B/16，
将编码的文本特征作为分类器来对每个文本进行分类
最后两层输出的稠密标记。为了使
嵌入到同一个特征空间中，我们让
倒数第二层输出的稠密 token 将其余部分传递
无自注意力层：

$$x_{密集} = x^{L-1} + c^{左}, \quad (6)$$

$$= x^{L-1} + x_{L-1}^{L}, \quad (7)$$

$$xdense = xdense + MLP(xdense)。 \quad (8)$$

我们在图 3 中提供了定性和定量结果
和表 1。结果表明，空间信息
倒数第二层保存完好，最后一层缺失。
因此，可以省略最后的自注意力操作，并根据投影输出进行分类

倒数第二层来发现目标类别的判别特征。

基于CLIP的多标签分类

本节介绍我们提出的局部到全局多标签分类框架，包括块级

分类以获得粗略分数，双掩蔽注意细化（DMAR）以细化粗略分数，以及

类别重新识别（CWR）模块进行双重检查
潜在的预测。

粗分类为了进行块级分类，基于倒数第二层的输出特征图

xdense ∈ RN×D被利用。文本编码器的输出
表示为 T ∈ RD×C，作为基于 ，
在文本输入上。N、D、C分别表示token长度、token维度和类别数。分类结果

xdense中每个补丁的分数计算如下：

$$si = 线性（xdense，i） \quad T， \quad (9)$$

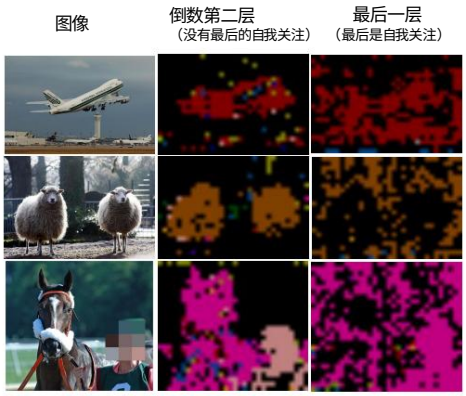

图 3：基于xdense和 x 的斑块级分类的定性结果
^{左}最后两层输出
CLIP-ViT 的最后一个自注意力操作
打破了 ViT 中的空间信息。我们模糊了人类
面临伦理方面的考虑。

最后的自注意力	mAP	mIoU
	82.7	16.2
	85.4	41.6

表 1：最后的自注意力操作在 PASCAL VOC 2012 验证集上对分类（mAP）和
分割（mIoU）的影响的定量结果。

其中 i 表示每个块的空间索引。线性
是 CLIP 的最后一层，用于映射编码图像特征
和文本特征到 CLIP 的统一空间中。si反映
图像标记和 C 语言文本描述的相似性，以及
相似度分数将被转发到softmax函数，以对所有类别的分数进行归一化（注意

softmax 操作是可选的，但我们发现它对于
CLIP 并在实验部分进行验证）。每个密集标记 i 的类别 c 的概率分类分数

可按如下方式获得：

$$P_{粗略}（i， c）= \frac{\exp(s_c)}{\sum_{k=1}^{碳} 指数（_{千}）}。 \quad (10)$$

双掩蔽注意力细化（DMAR）从公式 10 获得的初始块级分类分数经常受到噪声的影响，从而阻碍它们发挥作用

作为类别识别的可靠标准（例如，领先
图 5 中分类的假阳性）。先前的方法通常利用成对亲和力来细化密集

分类图，但需要训练额外的层（Ahn
and Kwak 2018；Ahn, Cho, and Kwak 2019）。相比之下，
Vision Transformer 固有的自注意力机制能够捕捉图像块之间的成对亲和力，使我们能够
改进块状分类分数，而不会产生
额外的计算成本。一种常见的方法是直接
使用最后几层的注意权重（Xu et al.
2022b) 或 ViT 的所有层 (Gao et al. 2021) 并执行

细化如下：

$$P_{精炼} = \frac{1}{|\psi|} \sum_{l \in \psi} To\ P_{粗}, \quad (11)$$

其中Pcoarse ∈ RN×C表示粗分数图， Al ∈
RN×N表示第 l 层注意力权重
ViT,ψ表示所用注意层的指标集，
|ψ| 是其元素数量。
然而,原始 ViT 中捕捉到的亲和力
MHSA 并不准确 (Ru et al. 2022),可能会误导改进过程。为了解决这个问题，
我们提出了一种双掩蔽策略,其核心思想是同时忽略注意力权重 A ∈
RN×N×L和

粗分数图Pcoarse∈RN × C。 对于注意力权重，
我们生成一个注意力掩码Mattn ∈RN ×N来选择
通过投票方式增加自信元素
在所有 L 个注意力层上。每个置信位置都应该
具有突出的关注价值（超过逐层平均值
值)至少在K层中,可以表示为：

$$Mattn(i, j) = 1,\text{如果} \sum_{l=1}^{左} I(A(i,j,l) > A^{-}l)\ (A) > K, \quad (12)$$

其中 I 是指示函数,A⁻ 。是平均值
第 l 层。细化过程如下：

$$P_{精炼} = \frac{1}{|\psi|} \sum_{l \in \psi} Mattn \odot Al\ P_{粗} \quad (13)$$

其中 ⊙ 表示 Hadamard 积。对于粗略得分
地图,我们根据计算每个类别的平均分数
P 通过忽略 并产生一个扩展的类掩码Mcls ∈
不自信的位置（低于平均得分)来改进RN×N×C 。每个类别c的最终改进分数可以是

得到如下结果：

$$Pref\ ined(c) = \frac{1}{|\psi|} \sum_{l \in \psi} Mattn \odot Al \odot Mcls(c)\ Pcoarse(c). \quad (14)$$

类别重新识别（CWR)虽然块级分类可以通过判别性局部特征发现目标类别，
但它可能会导致错误分类

缺乏全面的视角。因此,我们建议
一个类别重新识别模块,以进一步弥补
从全局视角对每个类别的主要预测分数。
具体来说,给定精细分类分数Prefined ∈
RN×C ,我们可以通过以下方式获得每个类的置信度Plocal
相应的最突出的补丁：

$$P_{局部}(c) = \max_{i} (Pcoarse(i, c)), \quad (15)$$

对于每个类别,我们挑选出反应灵敏的补丁
从Pref ined中形成类相关区域（类
蒙版）。我们通过区域的边界框裁剪图像
并将其调整为特定大小,例如 224×224。类级
mask 用作 ViT 中的注意掩码,以排除补丁
不属于该类别。我们输入类别图像
进入原始 CLIP 并使用 [cls] 标记进行分类。

获得的全局结果Pglobal与局部分数合并
Plocal利用本地和全局视图。

$$P_{最终} = \lambda Plocal + (1 - \lambda) Pglobal, \quad (16)$$

其中λ是平衡局部和全局效应的系数
在我们的实验中,它被简单地设置为 0.5。通过这个
融合过程中,我们可以有效地将宝贵的
结合本地和全球视角提供的见解,从而
提高整体分类性能。

下游任务的应用

多标签分类是一项实用任务,在依赖图像级标签的下游任务中有着广泛的应用。
本文探讨了 TagCLIP 的使用

结合现有的弱监督语义
分割（WSSS)方法用于解决无注释语义分割问题。给定图像级标签,大多数
WSSS
作品（Wang et al. 2020; Xie et al. 2022)杠杆类
激活映射（CAM)用于查找目标类的相关
图像中的区域并生成基于
就此而言。使用类别信息提供了宝贵的
高层指导,WSSS成绩斐然
甚至接近全监督的表现
设置。我们选择 CLIP-ES (Lin et al. 2023),因为它具有出色的准确性
和效率。它也是一种无需训练的
基于冻结 CLIP 的框架,更多详细信息请参见
发现于 (Lin et al. 2023)。通过利用这种高效的 WSSS
方法,整个分类然后分割范式需要
无需针对特定数据集进行训练,可以实现无注释
分割。我们将此框架称为 CLS-SEG。

实验

实验设置

数据集和评估指标。为了验证多标签分类的性能,为了进行公平的比较,我们

在 PASCAL VOC 2007 上评估我们的方法（Everingham
等人,2010)和 MS COCO 2014（Lin 等人,2014）。随后（Guo
等人,2023）。PASCAL VOC 2007 包含 20
类别,并使用 4952 张图像对测试集进行评估。
MS COCO 2014 包含 80 个类别,我们采用
官方拆分后,40137 张图像作为验证集。
对于下游语义分割,我们在三个常用数据集上进行实验,包括 PASCAL

VOC 2012（Everingham 等人,2010）,MS COCO 2017（Lin
et al. 2014) 和 COCO-Stuff (Caesar, Uijlings, and Ferrari 2018)。
对于 Pascal VOC 2012,有 20 个前景
其余像素为背景。包含 1449 幅图像的验证集用于验证。COCO

2017 年有 5000 张验证图像,包含 80 个类别和
背景类。COCO-stuff 有 4172 张验证图像
171个低级类别。我们采用27个中级类别,具体设置如下（Shin、Xie和Albanie，
2022b）。
请注意,我们的分类框架 TagCLIP 和分割框架 CLS-SEG 都是无需训练的,并且
可以直接在验证集上进行评估。我们采用均值
平均精度（mAP)作为多标签分类的评估指标,平均交并比

（mIoU)用于语义分割。

方法	额外训练数据	VOC	COCO
督导专科医生：			
南非储备银行	10% 数据	83.5	75.5
双重合作	10% 数据	90.3	78.7
或DPT	10% 数据	93.3	81.5
开放词汇通才：			
或DPT	COCO 字幕	88.3	65.1
夹子†	没有任何	79.5	54.2
夹子	没有任何	85.8	63.3
DPT†	没有任何	83.4	59.6
DPT	没有任何	86.2	64.3
CLIPSurgery	没有任何	85.4	61.2
TagCLIP（我们的）	没有任何	92.8	68.8

表 2:多标签分类的实验结果。 †
表示在分类分数上不使用softmax。

实施细节。我们的实验基于
ViT-B/16 已通过 CLIP 预训练。对于多标签分类，
图像保持原始分辨率。在每次操作中
其中需要置信度阈值，阈值 0.5 为
如果没有特别规定，例如阈值
在 CWR 中选择高响应度斑块。我们采用
CLIP 中使用的 80 个提示（Radford 等人，2021)和背景集（Lin 等人，2023）。
为了确定潜在的
根据分类逻辑对图像进行分类，我们首先
执行最小-最大正则化，将 logits 缩放到 [0, 1]
然后设置0.5来确定正类别。

实验结果

多标签分类。为了证明我们提出的 TagCLIP 的有效性，我们将其与其他

基于 CLIP 的方法。一些监督专家方法利用下游数据集的部分数据来训练

定制模型，包括 SARB（Pu 等人，2022 年）、Du-alCoOp（Sun、Hu
和 Saenko，2022 年）、TAI-DPT（Guo 等人，2022 年）
2023）。下游数据的使用限制了它们的泛化能力。另一种基于训练的方式无法
访问下游数据，而是使用精选的字幕数据进行训练，从而实现任意类别识别。其
他方法仅仅是

基于冻结的 CLIP，从而继承了其出色的泛化能力，包括 CLIP (Radford et al.
2021)，
DPT（Guo 等人，2023）、CLIPSurgery（Li 等人，2023）。
在表 2 中， † 表示直接将 softmax 之前的 logits 视为分类分数 (Guo
et al. 2023)，因为这些
logits 可以反映图像和文本特征之间的相似性。我们发现，性能下降显著

没有softmax激活，这可能源于使用
CLIP 预训练过程中的对比损失。结果
表 2 表明我们提出的框架表现
出奇地好。它增强了多标签分类
原始 CLIP 的性能大幅提升，即 7.0%
在 VOC 和 COCO 上分别提高了 5.5%。我们的方法超越了所有不需要额外训
练数据的方法。
VOC和COCO。它也与作品相媲美
需要额外的数据和训练。更多实验结果
请参阅附录。

方法	你	COCO	COCO 的东西
Vanilla USS 方法			
	9.8	-	6.7
蒙版对比度	35.0	3.73	-
特兰斯弗吉尼亚大学	37.2	12.7	17.5
面具蒸馏	45.8	-	-
喝	-	-	13.8
饮酒+H	-	-	14.4
基于CLIP的方法			
MaskCLIP ‡ 42.1 CLIPSurgery ‡ 41.5		20.2	23.9
GroupViT 52.3 SegCLIP 52.6 ReCo		25.2	29.7
34.2 NamedMask 59.2 CLS-SEG（我们		24.3	-
的)64.8 CLS-SEG （我们的)68.7		26.5	-
		17.1	26.3
		27.7	-
		34.0	30.1
		35.3	31.0

表 3:无注释语义分割的结果。
原始 USS 结果基于 K 均值聚类。 ‡
表示我们用相同的实验重新实现它
设置为我们的。 表示使用denseCRF进行后处理。

分割性能。我们提供无注释的分割结果，并使用 TagCLIP 生成的标签

作为伪标签并将它们与原始 USS进行比较
方法（包括 IIC（Ji、Henriques 和 Vedaldi 2019），
MaskContrast（Van Gansbeke 等人，2021）、TransFGU（Yin
等人。 2022），MaskDistill（Van Gansbeke、Vandenhende 和
Van Gool 2022)、PiCIE(+H) (Cho et al. 2021)) 以及最近
基于 CLIP 的作品（包括 MaskCLIP（Zhou、Loy 和
Dai 2022)、CLIPSurgery (Li et al. 2023)、GroupViT (Xu
等 2022a)、SegCLIP (Luo 等 2023)、ReCo (Shin, Xie,
和 Albanie 2022b），NamedMask（Shin、Xie 和 Albanie
2022a))见表 3。
我们观察到我们的 CLS-SEG 表现优于普通 USS
以及其他基于 CLIP 的方法，在三个方面都取得了很大的进步
数据集，这证明了我们生成的标签的高质量，并验证了这种先分类后分割范式的
有效性。从图4中，我们发现高级

类别信息提供的概念指导
图像对于获得高质量的分割掩模至关重要
因为：1)它可以防止因混淆而导致的错误预测
语义相似的类别中的纹理，例如皮肤
牛羊；2)可全面识别
一些类内方差较大的类别，例如，一个人的不同部分，可以通过更优的语义概念
被识别为一个整体。结果表明，分类有助于分割，并可能为以下领域提供启发

未来的研究。

消融研究

DMAR 和 CWR 的影响。在表 4 中，我们评估了
DMAR 和 CWR 在分类和
分割。DMAR 可以显著细化粗分，而 CWR 可以进一步提升性能。我们

在图5中提供了一个定性案例。DMAR之后，大多数不相关的类别可以被抑制。
CWR可以协调

图 4:MaskCLIP（Zhou、Loy 和 Dai 2022)和我们的方法的分割结果可视化。由于缺乏类别信息,MaskCLIP 的误报率更高。

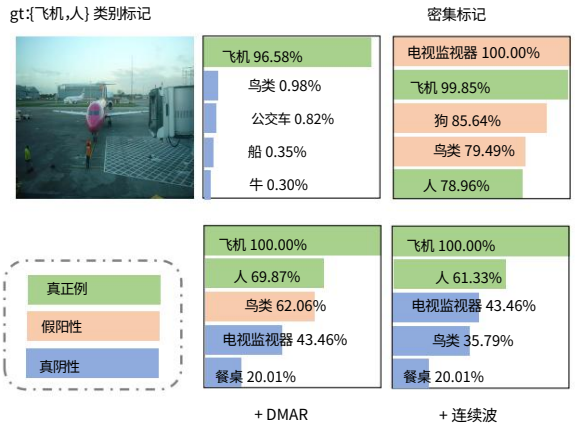

图 5:类标记和密集标记的分类结果以及我们提出的模块的效果。我们默认使用 0.5 作为阈值。

与DMAR相结合,从全局视角对精炼分数进行复核。因此,可以分别抑制和提高假阳性和假阴性的分数。

DMAR 模块中使用注意层的效果。为了确定 CLIP-ViT 中适合用于分类分数细化的注意层,我们首先比较了单层注意权重和多层注意权重在分类（准确率、召回率和 f1 分数)和分割（mIoU)性能方面的差异。从图 6 中,我们可以得出以下结论:1)融合多层注意权重通常比单层注意权重表现更好、更稳健。2)前几个注意层的性能不令人满意,这主要源于这些层学习到的注意力机制和特征较弱。3)最后一个注意层在最后几层中准确率较低,这与我们上面的分析相符。我们还展示了我们提出的双掩蔽策略的性能,该策略在大多数情况下有效地减轻了噪声的影响并改进了原始的注意力细化。该策略在仅略微降低召回率的情况下显著提高了准确率,从而总体上提高了分类和分割性能。

基于这些观察,我们在实验中融合了除最后一个之外的最后四个注意力权重。

粗略得分 DMAR CWR 图 mIoU		
	85.4 30.9	
	88.0 55.2	
	93.9 63.7	
	94.1	64.8

表4:DMAR和CWR模块在分类和语义分割方面的有效性结果。结果在PASCAL VOC 2012数据集上进行评估。

验证集。

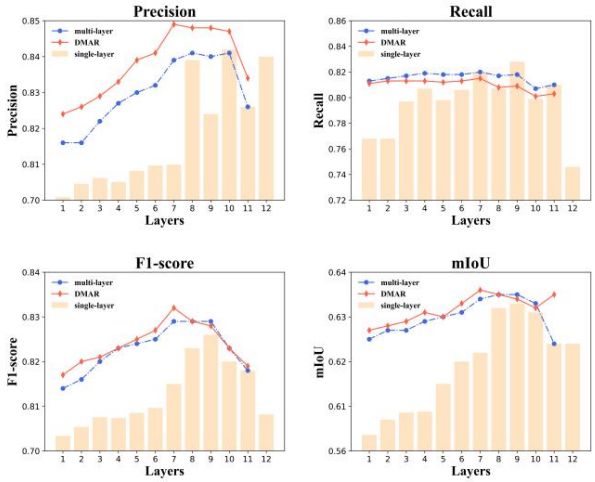

图 6:单层与多层注意力机制在分类和分割任务上的细化比较。对于单层设置,x 轴上的每个刻度 i 表示仅在第 i 层采用注意力权重。对于多层设置,第 i 个 x 刻度表示融合第 i 层至第 11 层的注意力权重,以细化粗分类分数。我们在融合过程中排除了最后一层注意力层。

结论

本文提出了一个简单有效的框架 TagCLIP,旨在增强原始 CLIP 模型的多标签分类能力。它遵循从局部到全局的范式,包含三个关键步骤:块级分类、双掩蔽注意力细化 (DMAR) 和类别重识别 (CWR)。得益于这些步骤,TagCLIP 释放了 CLIP 的潜力,并可作为通用的标注器,无需针对特定数据集进行训练即可提供高质量的图像标签。此外,我们验证了将生成的标签作为伪标签用于下游弱监督语义分割 (WSSS) 任务的可行性,并发现这种“先分类后分割”的范式显著优于以往自下而上的无标注分割方法。

这证明了 Tag-CLIP 的有效性和多功能性,并凸显了其在各种下游应用中的潜力。

致谢本研究得到国家自然科学
基金（批准号：62273303、62273301、62273302、62036009、61936006）、宁波市重点研发计划（批准号：2023Z231、2023Z229）、甬江人才引进计划（批准号：2022A-240-G）、浙江省重点研发计划（2023C01135）和国家重点研发计划（批准号：2022ZD0160100）的资助。

参考文献

Ahn, J.; Cho, S.; 和 Kwak, S. 2019. 具有像素间关系的实例分割的弱监督学习。
在 CVPR 中。

Ahn, J.;Kwak, S.,2018。通过图像级监督学习像素级语义亲和力,实现弱监督语义分割。载于 CVPR。

Ben-Cohen, A.;Zamir, N.;Ben-Baruch, E.;Friedman, I.;以及 Zelnik-Manor, L.,2021 年。零样本多标签分类的语义多样性学习。ICCV,第 640–650 页。

Caesar, H.;Uijlings, J.;以及 Ferrari, V.,2018 年。Coco-stuff：上下文中的事物和材料类别。CVPR,第 1209–1218 页。

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; 以及 Yuille, AL,2017。Deeplab：基于深度卷积网络、空洞卷积和全连接 crfs 的语义图像分割。TPAMI,40(4): 834–848。

Chen, T.;Xu, M.;Hui, X.;Wu, H.;以及 Lin, L. 2019a. 学习用于多标签图像识别的语义特定图表征。ICCV,522–531。

陈志梅;魏晓珊;王平;郭英;2019b。
基于图卷积网络的多标签图像识别。CVPR,5177–5186。

Cho, JH;Mall, U.;Bala, K.;以及 Hariharan, B.,2021 年。Picie：利用聚类中的不变性和等变性进行无监督语义分割。CVPR,第 16794–16804 页。

Crowson, K.;Biderman, S.;Cornis, D.;Stander, D.;Halla-han, E.;Castricato, L.;以及 Raff, E.,2022 年。Vqgan-clip：基于自然语言引导的开放域图像生成与编辑。载于 ECCV，第 88–105 页。Springer。

Everingham, M.; Van Gool, L.; Williams, CK; Winn, J.; 及 Zisserman, A. 2010. Pascal 视觉对象类 (Voc) 挑战赛。IJCV, 88(2): 303–338。

Gao, W.;Wan, F.;Pan, X.;Peng, Z.;Tian, Q.;Han, Z.;Zhou, B.;以及 Ye, Q. 2021. Ts-cam：用于弱监督物体定位的标记语义耦合注意力图。ICCV,第 2886–2895 页。

Ghiasi, A.; Kazemi, H.; Borgnia, E.; Reich, S.; Shu, M.; Goldblum, M.; Wilson, AG; 和 Goldstein, T. 2022。视觉转换器学到了什么?视觉探索。arXiv 预印本 arXiv:2212.06727。

Gu, X.;Lin, T.-Y.;Kuo, W.;以及 Cui, Y. 2021。通过视觉和语言知识提炼进行开放词汇对象检测。arXiv 预印本 arXiv:2104.13921。

Guo, Z.;Dong, B.;Ji, Z.;Bai, J.;Guo, Y.;和 Zuo, W.
2023. 将文本作为图像进行快速调整以实现多标签图像识别。CVPR,第 2808–2817 页。

He, S.; Guo, T.; Dai, T.; Qiao, R.; Shu, X.; Ren, B.; and Xia, S.-T. 2023. 通过多模态知识迁移实现开放词汇多标签分类。载于 AAAI,第 37 卷,808–816 页。

Huynh, D.;Elhamifar, E.,2020. 一种用于多标签零样本学习的共享多注意框架。CVPR，8776–8786。

Hwang, J.-J.;Yu, SX;Shi, J.;Collins, MD;Yang, T.-J.;Zhang, X.;以及 Chen, L.-C. 2019。Segsort：通过对片段进行判别排序实现分割。ICCV,7334–7344。

Ji, X.;Henriques, JF;以及 Vedaldi, A.,2019 年。用于无监督图像分类和分割的不变信息聚类。ICCV,9865–9874。

柯 T.-W.;黄建军;郭英;王晓燕;余绍雄
2022. 基于多视图共分割和聚类变换器的无监督分层语义分割。CVPR,2571–2581。

Li, Y.; Wang, H.; Duan, Y.; 和 Li, X. 2023。剪辑手术可在开放词汇任务中通过增强实现更好的可解释性。arXiv 预印本 arXiv:2304.05653。

Lin, T.-Y.;Maire, M.;Belongie, S.;Hays, J.;Perona, P.;Ra-manan, D.;Dollar, P.;以及 Zitnick, CL,2014。Microsoft coco：上下文中的常见对象。载于 ECCV。

Lin, Y.; Chen, M.; Wang, W.; Wu, B.; Li, K.; Lin, B.; Liu, H.; 以及 He, X. 2023. Clip 也是一种高效的分割器：一种基于文本的弱监督语义分割方法。CVPR,15305–15314。

Luo, H.;Bao, J.;Wu, Y.;He, X.;以及 Li, T.,2023 年。Seg-clip：基于可学习中心的块聚合方法,用于开放词汇语义分割。ICML,编号 23033–23044。

PMLR。

Narayan, S.;Gupta, A.;Khan, S.;Khan, FS;Shao, L.;以及 Shah, M.,2021 年。基于判别区域的多标签零样本学习。ICCV,8731–8740。

Pu, T.;Chen, T.;Wu, H.;以及 Lin, L.,2022 年。基于语义感知表征融合的多标签图像识别（部分标签）。载于 AAAI,第 36 卷,2091–2098 页。

Radford, A.;Kim, JW;Hallacy, C.;Ramesh, A.;Goh, G.;Agarwal, S.;Sastry, G.;Askell, A.;Mishkin, P.;Clark, J.;Krueger, G.;以及 Sutskever, I.,2021 年。《从自然语言监督中学习可迁移的视觉模型》。在 ICML 上发表。

Raghu, M.; Unterthiner, T.; Kornblith, S.; Zhang, C.; 以及 Dosovitskiy, A. 2021. 视觉转换器的视觉效果是否与卷积神经网络相似?NeurIPS, 34: 12116–12128。

Ren, S.;He, K.;Girshick, R.;以及 Sun, J.,2015 年。Faster r-cnn：基于区域提议网络实现实时目标检测。NeurIPS,28。

Ridnik, T.;Ben-Baruch, E.;Zamir, N.;Noy, A.;Friedman, I.;Protter, M.;以及 Zelnik-Manor, L.,2021 年。《多标签分类的非对称损失》。ICCV,第 82–91 页。

Ru, L.;Zhan, Y.;Yu, B.;以及 Du, B.,2022 年。通过注意力机制学习亲和力：基于 Transformer 的端到端弱监督语义分割。CVPR,第 16846–16855 页。

Selvaraju, RR;Cogswell, M.;Das, A.;Vedantam, R.;Parikh, D.;以及 Batra, D. 2017。
Grad-cam:通过基于梯度的定位从深度网络进行视觉解释。

在 ICCV 中。

Shin, G.;Xie, W.;以及 Albanie, S.,2022a。Namedmask:从互补基础模型中提取分割器。
arXiv 预印本 arXiv:2209.11228。

Shin, G.;Xie, W.;以及 Albanie, S.,2022b。ReCo:零样本迁移的检索与协同分段。刊于
NeurIPS。

Sun, X.; Hu, P.; 和 Saenko, K. 2022. Dualcoop:快速适应有限注释的多标签识别。

NeurIPS,35:30569–30582。

Van Gansbeke, W.;Vandenhende, S.;Georgoulis, S.;以及 Van Gool, L.,2021 年。通过
对比对象掩码提案实现无监督语义分割。ICCV,10052–10062。

Van Gansbeke, W.;Vandenhende, S.;以及 Van Gool, L. 2022. 使用 Transformer 发现
对象掩码以进行无监督语义分割。arXiv 预印本 arXiv:2206.06363。

Wang, X.;Wu, Z.;Lian, L.;Yu, SX,2022。从自然不平衡的伪标签中进行去偏差学习。CVPR，
14647–14657。

Wang, Y.;Zhang, J.;Kan, M.;Shan, S.;以及 Chen, X.
2020. 用于弱监督语义分割的自监督等变注意力机制。在 CVPR 上。

Wang, Z.;Chen, T.;Li, G.;Xu, R.;以及 Lin, L. 2017. 通过循环发现注意力区域实现多标签
图像识别。ICCV,第 464–472 页。

Wu, T.;Huang, Q.;Liu, Z.;Wang, Y.;以及 Lin, D. 2020 年。
长尾数据集中多标签分类的分布平衡损失。ECCV,162–178。Springer。

谢建军;侯晓玲;叶凯;沈玲,2022。CLIMS:用于弱监督语义分割的跨语言图像匹配。发表于
CVPR。

Xie, Q.;Luong, M.-T.;Hovy, E.;以及 Le, QV 2020。利用嘈杂学生进行自我训练可以改进
ImageNet 分类。
在 CVPR,10687–10698。

Xu, J.;De Mello, S.;Liu, S.;Byeon, W.;Breuel, T.;Kautz, J.;以及 Wang, X. 2022a.
GroupViT:语义分割源于文本监督。CVPR,18134–18144。

Xu, L.;Ouyang, W.;Bennamoun, M.;Boussaid, F.;以及 Xu, D. 2022b。用于弱监督语义分
割的多类标记转换器。发表于 CVPR。

叶建军;何建军;彭晓玲;吴伟;乔燕玲,2020 年。
用于多标签图像识别的注意力驱动动态图卷积网络。ECCV,649–665。Springer。

Yin, Z.;Wang, P.;Wang, F.;Xu, X.;Zhang, H.;Li, H.;以及 Jin, R.,2022 年。TransFGU:一
种自上而下的细粒度无监督语义分割方法。载于 ECCV,73–89。Springer 出版社。

You, R.;Guo, Z.;Cui, L.;Long, X.;Bao, Y.;以及 Wen, S.
2020. 基于语义图嵌入的跨模态注意力机制,用于多标签分类。载于 AAAI,第 34 卷,12709–
12716 页。

钟,Y。杨,J。张,P。李,C。科德拉,N.;李,L.H.;周L.;戴X.;袁L.;李,Y。等人。 2022. Region-clip：
基于区域的语言图像预训练。在 CVPR 中,16793–16803。

Zhou, B.; Khosla, A.; Lapedriza, A.; Olive, A.; 以及 Torralba, A. 2016. 用于判别定位的
深度学习特征。发表于 CVPR。

Zhou, C.;Loy, CC;以及 Dai, B.,2022。从剪辑中提取自由密集标签。载于 ECCV。

Zhou, K.; Yang, J.; Loy, CC; 和 Liu, Z. 2022. 学习提示视觉语言模型。国际计算机视觉杂
志,130(9): 2337–2348。

Ziegler, A.;以及 Asano, YM,2022。基于自监督学习的语义分割对象部分方法。CVPR，
14502–14511。

佐夫,B.;吉亚西,G.;林,T.-Y.;崔,Y。刘,H.;库布克,ED;和 Le, Q. 2020。重新思考预训练和自我
训练。神经IPS,33:3833–3845。

附录

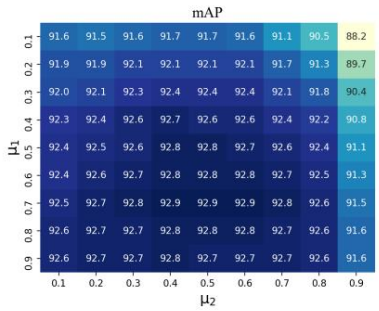

图 S1:在 VOC 2007 测试集上对μ1、 μ2的分析。

额外的实验

无缝线路参数的影响

在 CWR 模块中,引入了一些阈值来选择
潜在类别（μ1)和高度响应的补丁（μ2）。
我们在图 S1 中分析了μ1, μ2的影响。结果表明
分类性能对这些不敏感
参数具有适中的值。我们没有特别选择它们,只是在论文中简单地设置为0.5。

融合系数λ的影响

在我们的从本地到全球的框架中,我们融合了本地和
利用系数λ的全局分类得分，
控制局部和整体效应之间的平衡。
我们在表 S2 中分析了使用不同 λ 时分类性能的变化。结果表明,融合局部和全局特征可以提高分类性能,而单独使用局部特征（λ = 1)或全局特征则不能提高分类性能。

特征（λ= 0）,这验证了我们的
框架。我们没有特意选择最佳的λ,而是在实验中简单地将其设置为0.5。

背景设置效果

我们对以对象为中心的 PASCAL 进行了实验
VOC（Everingham 等人,2010）、MS COCO（Lin 等人,2014）
以及以场景为中心的 COCO-Stuff (Caesar、Uijlings 和 Ferrari 2018)
基准。然而,以对象为中心的数据集
仅包含特定对象的注释,其余部分被视为背景,这不利于

我们的块级分类。我们遵循 CLIP-ES (Lin et al.
2023) 来利用一些常用的背景类别。具体来说,PASCAL 中使用的背景集

VOC 是{地面、土地、草地、树木、建筑物、墙壁、天空、湖泊、
水,河,海,铁路,铁路,头盔,云,房子,
山、海洋、道路、岩石、街道、山谷、桥梁、标志、键盘}。对于 COCO,最后两个背景类别是
删除,因为在
数据集。COCO-Stuff 没有使用背景集,因为
定义的 171 个类别涵盖了事物和材料

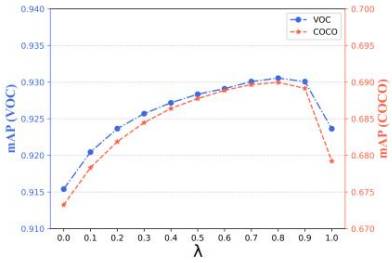

图 S2:λ 对分类性能的影响
PASCAL VOC 2007 测试集和 COCO 2014 验证集。

背景设置 VOC COCO		
	92.4	67.9
	92.8	68.8

表 S1:背景集对多标签分类任务的影响结果（mAP）。结果基于

PASCAL VOC 2007 测试集和 COCO 2014 验证集。

类别。我们展示了背景设置对
表S1中的分类和分割任务。它对整体表现有所贡献。

限制

提出的 TagCLIP 遵循从本地到全局的框架
增强 CLIP 的多标签分类性能。它在常见的多标签数据集上表现出色

（例如 VOC 和 COCO）,其中定义的类别是相互竞争的,例如猫和狗。然而,
TagCLIP
对于具有包含关系的数据集可能不是最佳选择
类别之间,例如猫和动物。这是因为
层次式类别不太适合预先训练的
CLIP 算法基于对比损失,并促进不同类别之间的竞争。一个潜在的解决方案是采用自训练方法。此外,某些场景类别（例如天空)可能无法从判别性局部线索中显著获益,需要更强的条件才能实现准确识别。因此,我们的方法在以物体为中心的数据集上表现优异,但仍有提升空间。

面对不同粒度的层次类别时,改进方法至关重要。我们将解决这个问题作为未来的研究方向。

额外的定性结果

在图 S3 中,我们利用我们的
CLS-SEG 框架在不同基准上的表现。我们的方法
在这些具有挑战性的数据集上表现良好,尤其是对于
刚性物体和动物。稀有类别和物品类别仍有改进空间,这可能是

通过更精确的类别描述来解决。

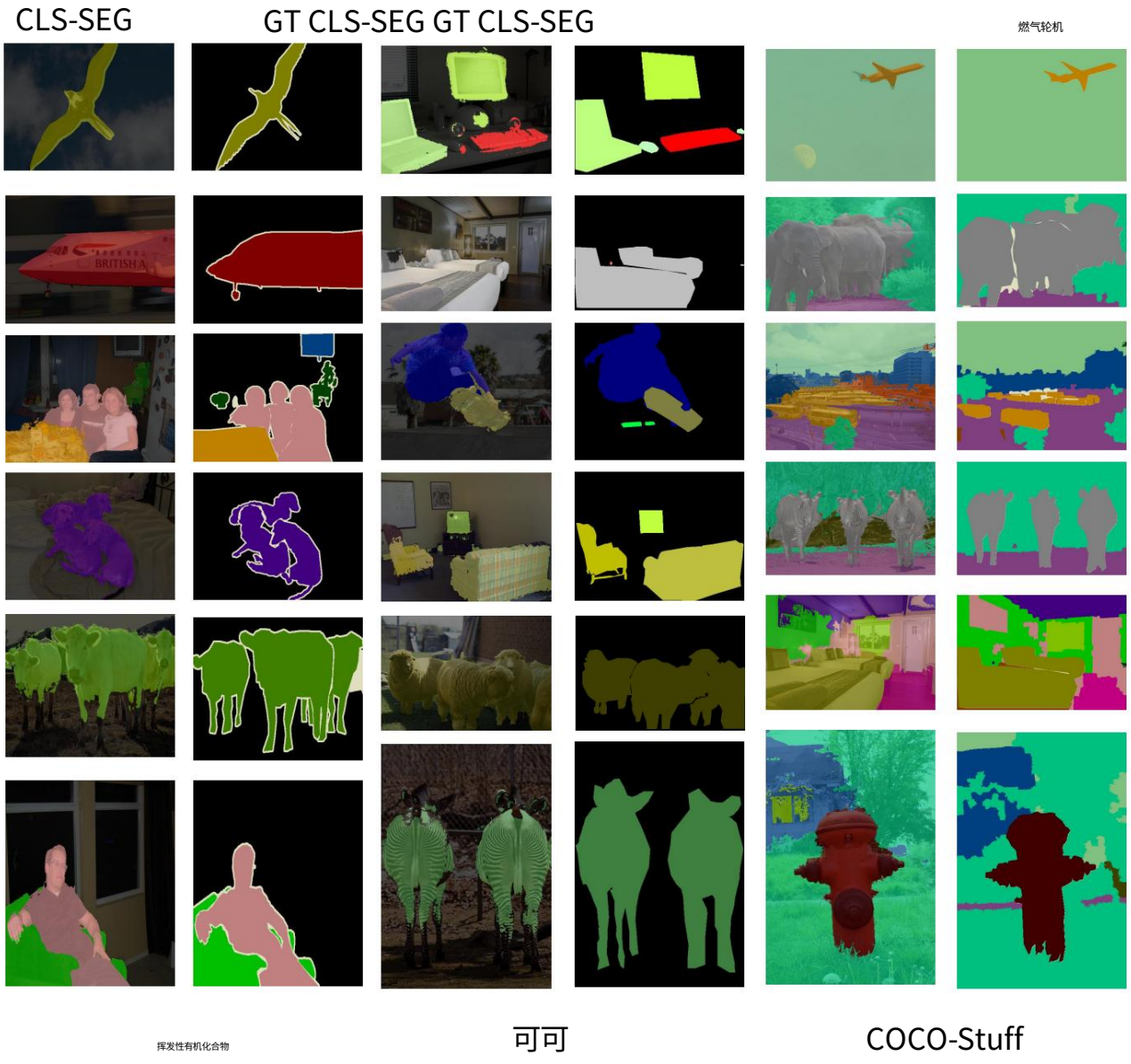

图 S3:我们的方法在 VOC、COCO 和 COCO-Stuff 上的附加可视化。