迷失在翻译中？源自由域适应中的词汇对齐在开放词汇语义分割中的应用

Silvio Mazzucco[1,2*]
silvio.mazzucco@thegoodailab.org
Carl Persson[1*]
carl.persson@thegoodailab.org
Mattia Segu[1,2,3]
mattia.segu@thegoodailab.org
Pier Luigi Dovesi[1]
pier@thegoodailab.org
Federico Tombari[3,4]
tombari@google.com
Luc Van Gool[5]
vangool@vision.ee.ethz.ch
Matteo Poggi[1,6]
m.poggi@unibo.it

[1] The Good AI Lab
the good ai lab 组织
[2] ETH Zurich
[3] Google
[4] Technical University of Munich
[5] INSAIT, Sofia University, St. Kliment Ohridski
[6] University of Bologna

* joint first authorship

摘要

我们介绍了一个全新的无源域适应框架**声学对齐**，专门用于开放词汇语义分割中的 VLMs。我们的方法采用了一种增强的师生范式，结合了词汇对齐策略，通过引入额外的类概念来改进伪标签生成。为了确保效率，我们使用低秩适应（LoRA）对模型进行微调，在保持其原始能力的同时最小化计算开销。此外，我们为学生模型提出了一种 *Top-K* 类选择机制，这显著减少了内存需求并进一步提高了适应性能。我们的方法在 CityScapes 数据集上取得了显著的**+6.11 平均交并比**提升，并在零样本分割基准测试中表现出色，为开放词汇设置下的无源域适应设定了新标准。

项目页面：https://thegoodailab.org/blog/vocalign

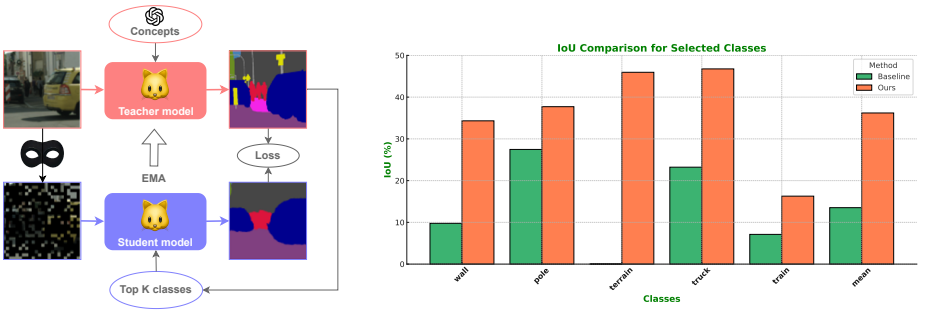

图 1: **方法概述**。左：我们的师生框架及其引入的额外技术。右：VocAlign 对 CityScapes 中选定类别的影响。

1 介绍

开放词汇语义分割旨在为图像中的每个像素分配一个类别标签，通过移除固定类别的约束来扩展传统的语义分割。这得益于视觉-语言模型（VLM），它们利用多模态协同作用并基于网络规模的数据集进行预训练。虽然这种能力允许更广泛的泛化，但也使得 VLM 对领域迁移更加敏感，在更为复杂的联合视觉文本空间中发生的这些变化往往导致在不同类别未见数据集上的表现不佳。

无监督领域适应（UDA）已被广泛研究作为解决领域偏移 [45] 的一种解决方案，允许模型在无需目标领域的标注情况下适应新领域。然而，大多数 UDA 方法 [1, 18, 19, 20] 假设可以访问原始模型训练所用的源数据。这一假设对于通常在不公开的专有网络规模数据集上进行训练的视觉语言模型（VLMs）来说是不切实际的。

无源域适应（SFDA）通过允许模型在无需访问源数据的情况下进行适应来解决这一限制。虽然 SFDA 提供了一种更高效和可扩展的范式，但由于适应过程中缺乏标注样本，它也带来了挑战。鉴于无法访问其庞大的训练数据集是不现实的，这种方法特别适合于调整 VLMs。然而，现有的 SFDA 方法主要针对没有开放词汇能力或多模态视觉语言交互的模型设计，限制了它们在 VLMs 中的适用性。此外，VLM 训练数据集的巨大规模和多样性加剧了适应挑战，因为特征空间中的分布变化和标签重叠现象变得更加明显。

在这项工作中，我们提出了**声学对齐**，这是首个专门针对开放词汇语义分割的 VLMs 设计的 SFDA 框架。为了应对 VLMs 带来的独特挑战，我们采用了一种师生框架，并根据其多模态和开放词汇特性扩展了该框架，如图 1（左）所示。首先，通过增强教师模型词典中的额外类别概念来改进伪标签生成。这利用了 VLMs 的开放词汇能力，改善了视觉嵌入与伪标签在各领域间的对齐，从而实现

了更有效的适应。其次，为了管理大规模 VLMs 适应过程中的计算成本，我们采用低秩适配（LoRA）模块 [21] 进行参数高效的微调。该方法保留了从网络规模数据中学习到的原始知识，并最大限度地减少了额外开销。第三，我们在学生模型中引入了一种 *Top-K* 类别选择机制，在每次迭代中仅优化基于伪标签预测的一组类别的子集。这显著降低了内存需求而不会牺牲性能。

我们在 CAT-Seg[7] 上评估了我们的方法，这是一个最先进的开放词汇分割模型，使用 CityScapes 数据集。我们的方法在多个类别上取得了显著改进，包括完全恢复之前无法识别的类别，如图 1（右）所示。此外，我们将评估扩展到零样本分割数据集，展示了我们的方法在具有不同目标类数量的各种设置中的有效性。

我们的主要贡献总结如下：

- 我们提出了首个专门针对基于视觉语言模型的开放词汇语义分割模型的 SFDA 框架。
- 我们引入了一种词汇对齐策略，该策略通过利用 VLM 的多模态能力来提高伪标签的质量。
- 我们通过结合 LoRA 模块和 *Top-K* 类选择，降低了计算复杂度，在提高内存效率的同时进一步提升了性能。

2 相关工作

我们简要回顾了与我们的工作相关的文献，涵盖了开放词汇语义分割、无监督领域适应和参数高效微调方法。

基于视觉语言模型的语义分割。语义分割是受益于视觉语言模型（VLM）如 CLIP[15, 23, 43] 及其变体 [37] 出现的关键任务之一 [35, 56]。周等人 [55] 表明，除了从 CLIP 中提取全局特征表示外，还可以直接获得低分辨率分割图，而吴等人 [48] 则完善了提取和处理管道。

这些方法的共同目标是利用 CLIP 的知识来处理未见类别。Li 等人 [26] 引入了一个框架，在该框架中，文本嵌入与视觉表示显式对齐，最大化相关文本和视觉概念之间的关联性。类似地，OVSeg[29] 将 CLIP 主干适应以提高掩码提案机制的性能。最近的工作，如 [51, 59]，通过转向单阶段方法简化了流程。Cho 等人 [7] 提出聚合由 CLIP 生成的空间和类别维度上的成本体积，而 Xie 等人 [49] 将这一想法扩展以进一步提高准确性。

无监督领域适应。为下游任务（如图像分类 [12, 31, 32, 41]、对象检测 [5, 6, 28, 39] 和语义分割 [3, 16, 46, 60] 等）开发了多种 UDA 方法。传统的 UDA 方法

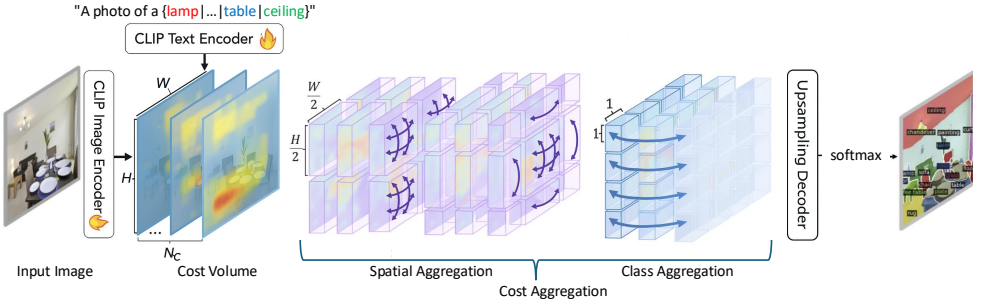

图 2: **CAT-Seg 主干网络**。在核心部分，成本聚合模块在空间和类别级别上发挥作用。最后，通过 softmax 层选择预测的类别。

[4, 18, 19, 20, 50, 52] 在适应过程中通常假设有源数据的访问权限。然而，由于隐私和可访问性问题，这种假设变得越来越不现实。因此，UDA 研究扩展到了无源和测试时适应设置中，这些设置明确解决了在适应期间没有源数据的问题。

在测试时适应（例如，[44]），额外的挑战在于在线调整模型，以考虑数据中分布变化的时间演变。刘等人。[30] 使用对抗训练，通过生成类似于源域数据的人工样本来引导适应。昆杜等人。[24] 通过对训练期间使用源数据然后将无源适应应用于目标域的方法来解决领域泛化问题。王等人。[47] 提出了一种具有增强平均伪标签的学生-教师框架。

最近，针对视觉语言模型的方法出现了。Samadh 等人 [40] 利用来自视觉和文本编码器的可学习提示在适应过程中的对齐。Choe 等人 [8] 探索了语义分割的开放集领域自适应，尽管不是在无源设置中。

参数高效微调方法。这些方法通过减少可学习参数的数量来应对适应数十亿参数模型的挑战，同时充分利用预训练权重的全部潜力。

最流行的其中一种方法涉及低秩变换。受到 [21] 的启发，这些方法通过将更新的权重增量分解为两个低秩矩阵来进行重新参数化。在此基础上，诸如 [53] 等方法探索动态调整 LoRA 矩阵的秩，而 [22] 则研究结合多个 LoRA 模块以增强灵活性和性能。

作为一种替代方案，基于适配器的方法已被广泛用于参数高效的微调 [17, 36]，而基于提示的调整 [25, 27] 则将提示转化为可学习的参数以最小的开销提高视觉语言模型的准确性 [9, 13, 57, 58]。这些参数高效策略通常应用于 UDA[11, 14] 的背景下。

3 方法

我们现在介绍我们的框架，VocAlign。首先，我们介绍一些关于开放词汇语义分割和基线模型的背景知识，然后描述我们的教师-学生框架以及特定技术，使其适应开放词汇设置。

3.1 预备知识：CAT-Seg 主干网络结构

VocAlign 应用于 CAT-Seg[7]，这是一种最先进的开放词汇分割模型，如图 2 所示，该模型由三个组件组成：特征提取器（带额外视觉编码器的 CLIP 编码器）、成本聚合模块和上采样解码器。

特征提取器包括一个略微修改版的 CLIP 图像编码器以及标准的 CLIP 文本编码器。通过这一主干结构，CAT-Seg 抽取文本和密集视觉特征 $\mathcal{D}^L = \phi^L(t), \mathcal{D}^V = \phi^V(x^T)$，用于构建由多模态特征之间的余弦相似性 [38] 组成的成本体积 $\mathcal{C} \in \mathbb{R}^{H \times W \times P \times N_c}$

$$\mathcal{C}(i,n) = \frac{\mathcal{D}^V(i) \cdot \mathcal{D}^L(n)}{\| \mathcal{D}^V(i) \| \| \mathcal{D}^L(n) \|}, \tag{1}$$

其中 i 表示像素坐标，而 n 对应于 N_c 类别之一的文本嵌入。文本嵌入通过 P 种多样的提示进行丰富，例如“一幅类的画”或“一个类的渲染”，从而得到形状为 $\phi^L \in \mathbb{R}^{N_c \times P \times d_L}$ 的嵌入。

为了细化粗糙的成本体积，CAT-Seg 使用一个具有两种不同聚合机制的成本聚合模块，这两种机制均通过 Swin Transformers 实现：i) 空间聚合，建模像素级空间交互以传播信息；和 ii) 类聚合，专注于不同语义类别之间的交互。

细化后，解码器将体积上采样以匹配输入分辨率。然后将输出传递给 softmax 层以生成最终预测。

3.2 学生-教师框架

作为 VocAlign 的基础，我们采用学生-教师知识蒸馏 [42] 将知识从源域转移到目标域。最初，在源域上使用源图像 $\mathcal{X}^S = \{x_k^s\}_{k=1}^{N_s}$ 及其相应标签 $\mathcal{Y}^S = \{y_k^s\}_{k=1}^{N_s}$ 训练一个神经网络 f_θ。在语义分割的背景下，最常见的损失是像素级交叉熵：

$$\mathcal{L}_k^{S,seg} = \mathcal{H}(f_\theta(x_k^S), y_k^S), \qquad \text{with} \qquad \mathcal{H}(\hat{y}, y) = -\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{c=1}^{C} y_{ijc} \log \hat{y}_{ijc} \tag{2}$$

在我们目标的 SFDA 设置中，部署后源图像和标签将不再可用。从这一点开始，训练数据指的是缺乏标签的目标图像 $\mathcal{X}^T = \{x_k^T\}_{k=1}^{N_T}$。

预训练模型作为我们 SFDA 框架中的教师 g_ϕ，生成伪标签来监督另一个模型实例，即学生 f_θ，该实例被调整以适应目标领域。伪标签是通过将目标图像输入到 g_ϕ 中生成的。

$$p^T_{i,j} = [c = \arg\max_{c'} g_\phi(x^T)_{i,j,c'}], \tag{3}$$

其中 $[\cdot]$ 表示 Iverson 括号。我们还通过利用由 g_ϕ 预测的类概率来加权损失，计算置信度超过阈值 τ 的像素比例。置信度 q^T 用每个像素的最大 softmax 概率表示：

$$q^T_{i,j} = \frac{[\sum_{i=1}^{H}\sum_{j=1}^{W} \max_{c'} g_\phi(x^T)_{i,j,c'} > \tau]}{H \cdot W} \tag{4}$$

在适应过程中，教师模型保持冻结，没有梯度回传通过它。然而，它是随着时间作为学生权重 [42] 的指数移动平均（EMA）进行更新的，具体为：$\phi_{t+1} \leftarrow \alpha\phi_t + (1-\alpha)\theta_t$，其中 $\alpha \in [0,1]$ 控制学生更新对学生师模型的影响程度。

学生输入掩码。遵循 [20]，我们将掩码应用于输入到 f_θ 的图像以提高适应性。二进制掩码的采样方式为：

$$\mathcal{M}_{\substack{mb+1:(m+1)b \\ nb+1:(n+1)b}} = [v > r] \quad \text{with} \quad v \sim \mathcal{U}(0,1), \tag{5}$$

其中 $[\cdot]$ 是伊弗森括号，b 是块大小，r 是掩码比例，$m \in [0,...,W/b-1]$，$n \in [0,...,W/b-1]$ 是块索引，$\mathcal{U}(0,1)$ 是均匀分布。学生预测 $\hat{y}^{\mathcal{M}}$ 完全基于被遮罩图像的可见上下文：

$$\hat{y}^{\mathcal{M}} = f_\theta(x^{\mathcal{M}}) \qquad \text{with} \qquad x^{\mathcal{M}} = \mathcal{M} \odot x^T \tag{6}$$

最后，学生模型通过其预测与教师伪标签之间的交叉熵损失 $\mathcal{H}$ 进行监督：$\mathcal{L}^{\mathcal{M}} = q^T \mathcal{H}(\hat{y}^{\mathcal{M}}, p^T)$，其中 p^T 是伪标签，q^T 是基于置信度的权重。

3.3 词汇对齐

为了应对开放词汇设置中的挑战，VocAlign 在文本空间中引入数据增强，而不是仅仅依赖图像增强 [47]。我们通过 概念——额外的文本描述或同义词来丰富目标数据集类别，这些仅在适配过程中使用。由于源模型是在大规模数据集上预训练的，它倾向于其训练数据中存在的类别，这些类别可能只部分重叠或以不同的名称存在于目标类别中。这是 VLMs 在 SFDA 中的新问题，因为领域转移不仅涉及视觉外观，还涉及视觉-文本对齐。

例如，如果在预训练期间存在 “动物” 类，将其作为概念添加到目标类 “猫” 中可以提高性能，特别是在目标数据集中不存在其他与动物相关的类别时。相应地，成本体积扩展为 n_{tot} 类别，包括原有的类别加上概念。每个类别 n，$n = 1,\ldots,N_c$

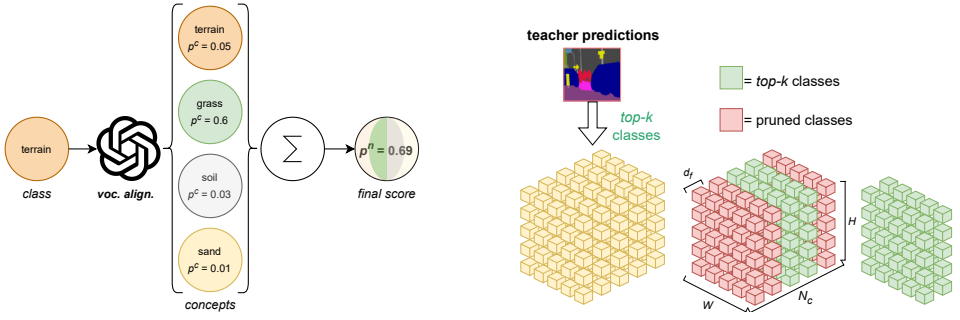

图 3: **VocAlign 核心组件**。左边：CityScapes 上的词汇对齐示例。右边：Top-K 选择的实际操作。

可以被增加 $c(n)$ 个概念。在生成伪标签之前，概念会被重新聚合到它们原来的类别中：

$$p_{ij}^{c} = \frac{e^{x_{ij}^{c}}}{\sum_{c'=1}^{n_{\text{tot}}} e^{x_{ij}^{c'}}}, \quad c = 1, \ldots, n_{tot}, \qquad p_{ij}^{n} = \sum_{h=1}^{C(n)} p_{ij}^{c_h(n)}, \quad n = 1, \ldots, N_c \tag{7}$$

其中 p_{ij}^{c} 是涵盖概念的 n_{tot} 类别的概率，而 p_{ij}^{n} 则是 N_c 原始类别的最终概率。示例如图 3（左）所示。

为了简化大型数据集的概念创建过程，我们使用 ChatGPT o1 自动生成这些概念，设计的提示语能够捕捉上下文、目标和数据集信息。更多细节见补充材料。

3.4 Top-K 选择

适应许多类可能是计算上不可行的。为了解决这个问题，我们使用教师预测（增强的概念）来识别图像中的可能类别。给定教师概率 $\hat{y} \in \mathbb{R}^{N_c,H,W}$，我们计算每个类别的平均激活值，并选择具有最高平均值的 *Top-K*。剩余的类别在聚合之前从成本体积中被剪枝，如图 3（右）所示。

因此，每次迭代中只有部分类别接受监督。尽管这似乎减少了监督，但教师生成的高质量伪标签确保了现有类别的可靠预测，最终为这些类别提供了更强的监督。

4 实验

我们现在介绍我们的实验结果。首先，我们描述实现细节、评估中使用的数据集和训练调度，然后报告我们的主要实验，并以一些分析和消融研究作为结论。

4.1 实现细节

我们的方法基于毫米分割框架 [10] 及其 CAT-Seg 实现，我们构建了一个学生-教师框架，集成了 [20] 提供的代码。我们使用了带有 Resnet-101 和 ViT-B 编码器的 CAT-Seg 主干网络，并对其进行了修改，以使用重复八次的 10 个提示模板，而不是原始模型中的 80 个模板，以提高效率——我们还将提供使用完整 80 个模板的结果。我们从在 COCO-Stuff[2] 上预训练的 CAT-Seg 权重开始，并在我们报告的任何实验中调整这些权重。我们仅在 CLIP 主干网络中引入 LoRA，针对图像编码器的 ViT 的前 4 层和文本编码器的 Transformer 的注意力投影矩阵。在标准设置中，我们使用等于 2 的 LoRA 秩。我们保持模型的其余部分冻结。

数据集。我们主要使用 CityScapes 进行训练和评估，因为它是一个常用的用于评估领域适应方法的数据集，并且它与 COCO-Stuff 足够不同以用作一个有利的目标域。它包含 2975 张训练图像和 500 个验证样本，总共有 19 类。为了证明我们的方法也可以推广到其他数据集，我们使用了 ADE20K-150[54] 和 PASCAL-Context 59[34]。这些数据集被用来衡量原始 CAT-Seg 模型的零样本性能。ADE20k-150 包含 20k 张训练图像和 2k 张验证图像，总共有 150 类标注。PASCAL-Context 59 包含 5k 张训练和验证图像，带有 59 个类别标签。平均交并比（mIoU）被用作主要指标。

城市景观训练。我们使用 AdamW 优化器训练我们的方法，总共进行 40k 次迭代 [33]，在训练开始时经过一个快速预热阶段后学习率设为 5×10^{-5}，从 5×10^{-6} 开始进行了 500 次迭代。我们使用大小为 2 的批次和 Top-K 选择值为 15。输入数据通过随机裁剪图像（512×512 裁剪）以及应用随机颜色抖动来进行增强。图片的掩码比率为 0.7。与大多数方法一样，我们将 $\alpha = 0.99$ 设定为教师模型 EMA 更新的值。根据 CityScapes 类别描述手动选择了合适的概念。

多数据集训练。我们选择一个配置来应用于所有三个数据集以评估我们方法的泛化能力。我们训练了 15k 次迭代，并在适用时使用批处理大小为 2，否则则应用批处理大小为 1。输入图像通过随机裁剪到 512×512 个裁剪区域并应用随机翻转、光度失真和颜色抖动来进行数据增强。学习率选择为 3×10^{-5}，并且我们应用了 0.5 的掩码比率。所有数据集的概念均使用 ChatGPT-o1 生成。我们在训练修改后的包含 10 个提示模板的 CAT-Seg 模型时利用 Top-K 值为 50，在整个 CAT-Seg 模型上训练时则使用 Top-K 值为 35。

4.2 主要结果

城市景观的结果。我们在 CityScapes 上的初步结果如表 1 所示，这是由 VocAlign 改进后的前四优和最差类别（完整结果见补充材料）。VocAlign 展现了

Method	car	fence	road	mbike	...	pole	truck	wall	terrain	mIoU
Zero-Shot CAT-Seg	**76.38**	**38.37**	**86.03**	**55.35**	...	27.47	23.22	9.77	0.02	47.56
VocAlign (ours)	67.48	31.70	84.87	54.83	...	37.71	**46.76**	**34 .32**	**45.95**	**53.67**
提高 (%)	-8.9	-6.67	-1.16	-0.52	...	+10.24	+23.54	+24.55	+45.93	+6.11

表 1: **各类别中表现最好和最差的前四类在 CityScapes 数据集上的结果**。我们将 VocAlign 与 CAT-Seg 的零样本预测进行了比较。最后一列显示了所有类别的 mIoU。

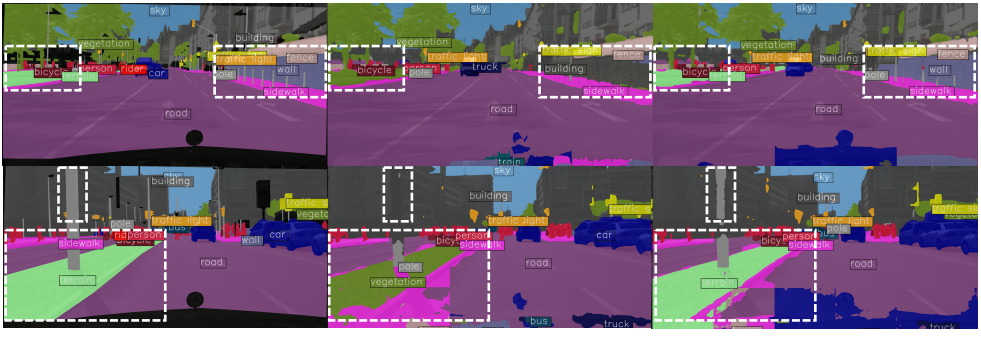

图 4: **CityScapes 数据集的定性结果**。从左到右：真实分割图，零样本 CAT-Seg 预测和 VocAlign 预测。

强大的适应能力，尤其是在零样本设置下表现不佳的类别中尤为突出。这包括地形类别，其 mIoU 评估值从接近零大幅提高。许多显著的改进被认为源于类别的名称和描述中的歧义。例如，在 CityScapes 中定义的墙类别为一个单独的墙体，不属于建筑物的一部分。然而，由于传递给分割模型的只有墙这个词，因此这个额外的类别信息丢失了。我们可以通过添加概念或修改教师类来恢复这一点，这很可能是此类别改进显著的原因。此外，我们发现我们的方法降低了某些类别的性能——最明显的是汽车类别。这被认为是由采集车辆的引擎盖引起的，在训练过程中没有对其进行屏蔽。模型因此学会了将一般区域以及一定程度上的附近路面分类为汽车。在验证期间屏蔽了引擎盖，导致部分道路被误分类为汽车，减少了该特定类别的得分，如图 4 所示。我们在补充材料中进一步证实了这一假设。

多数据集结果。我们在多个数据集上的评估结果如表 2 所示。相应地，我们的方法可以使用通用配置很好地推广到其他数据集，其中模型的适应性仍然很强。与 CityScapes 训练相比，性能下降部分原因在于使用了由 ChatGPT 生成的提示以及所有数据集中迭代次数较少。该方法还应用于完整的 CAT-Seg 模型（右侧），我们继续展示中等程度的适应性。评估得分较低的原因可能是因为受限于我们的内存限制——单个 64GB A100 GPU，导致 Top-K 选择值较低。

Method	CityScapes	PC-59	ADE20k-150
Zero-Shot	47.56	55.52	26.88
Min-Entropy	47.67	55.87	27.23
Teacher-Student	44.58	55.20	26.55
+ Masking	47.71	55.89	26.84
+ Vocab Alignment	49.58	56.81	26.77
+ TopK	**49.58**	**57.01**	**27.39**
提高 (%)	+2.02	+1.49	+0.51

Method	CityScapes	PC-59	ADE20k-150
Zero-Shot	47.88	56.94	27.22
Min-Entropy	47.95	-	-
VogAlign (ours)	**48.97**	**57.32**	**27.40**
提高 (%)	+1.09	+0.38	+0.18

表 2: **多数据集评估**。左边：10 提示 CAT-分割，最小熵基线和 VocAlign 的简化版本。右边：原始 80 提示 CAT-分割。

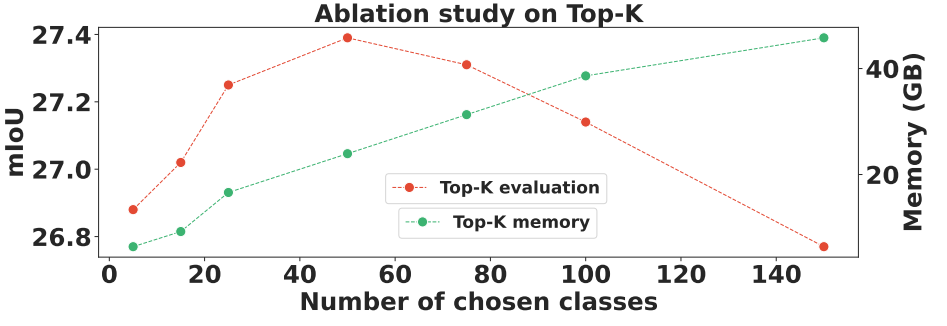

图 5: **ADE20k-150 的消融研究——Top-K 类别数量的影响**。我们改变由 Top-K 方法选择的类别数量。

与其他方法的比较。根据我们的知识，之前没有关于开放词汇语义分割模型的 SFDA 的相关研究。因此，我们实现了一个简单的基线方法，利用最小化熵作为训练目标。这个目标使得可以在无源方式下调整 CAT-Seg 模型。该方法的结果可以参见表 2。尽管与 VocAlign 相比，最小化熵的目标仅能在有限的程度上适应 CAT-Seg 模型。

4.3 分析与消融实验

模型组件分析。表 2（左）显示了每个模型组件的有效性。基准的教师-学生设置表现不佳，无法有意义地将模型适应于任何数据集。向模型中添加掩码显著提高了基线性能，但与 CAT-Seg 相比，增幅仅是适度或不存在的，这取决于数据集。我们的词汇对齐方法能够在 CityScapes 和 PASCAL-Context 59 上显著调整模型，然而我们注意到，在 ADE20K-150 中单独使用掩码时性能有所下降。这可能是由于数据集中类别数量庞大，难以找到合适的概念所致。最后，结合 Top-K 方法适度提高了任何数据集上的模型性能。

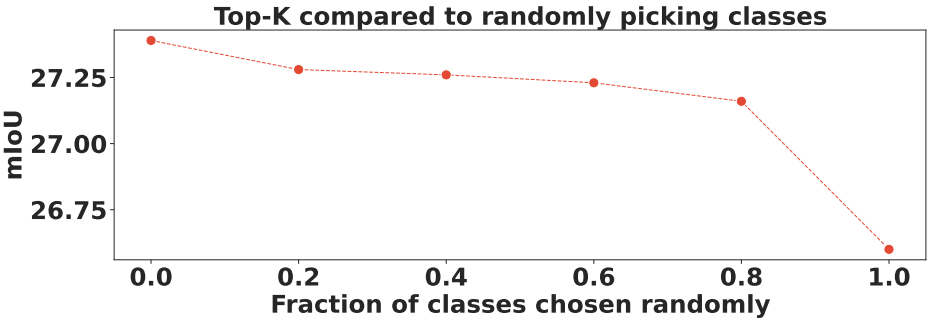

图 6: **ADE20k-150 的消融研究——Top-K 选择与随机选择对比**。我们用随机采样的类别替换一部分 Top-K 类别。

最终模型性能。表 2（右）表明我们的方法可以应用于完整的 CAT-Seg 模型，利用所有 80 个提示模板。然而，由于更高的内存需求，选择了一个较低的 Top-K 值为 35，以适应单个 GPU。这导致了性能有适度的提升。尽管如此，我们仍然发现在 CityScapes 上有很大的改进，在 PC-59 和 ADE20k-150 上有适度的改进，相比于零样本设置。

Top-K 选择的影响。图 5 和 6 证明了 Top-K 选择对 ADE20k-150 的影响。Top-K 最初作为一种节省成本的方法实现 – 如图 5 中的绿色曲线所示。然而，这种消融实验也揭示了它带来了适度的性能提升 – 见红色曲线。此外，我们在图 6 中展示了用随机类替换一部分 Top-K 选择的类别会导致更差的表现，证实了 Top-K 选择严格优于随机选取类别。使用 Top-K 带来的性能提高表明，最相关的类别接受了相对更强的梯度，而被剪枝的类别没有接收到错误的监督。

5 结论

我们介绍了 VocAlign，这是首个面向开放词汇语义分割模型的 SFDA 方法。这一方法通过伪标签生成的词汇对齐策略、学生模型上的顶部-K 类选择机制以减少内存需求以及巧妙使用 LoRAs 来实现。我们的方法在 CityScapes 数据集上取得了很好的结果，我们认为这是主要基准测试之一，因为它是在 UDA 中最常用的数据集之一。此外，我们在其他开放词汇数据集上也展示了有希望的结果，并以此证明解决了与 VLM 特性相关的问题。这项工作为该主题的进一步研究开辟了道路，未来这可能会变得更加重要。

致谢。我们感谢欧洲高性能计算联合事业（EuroHPC JU）、EuroCC 国家能力中心瑞典分部（ENCCS）以及在 ISCRA 倡议下 CINECA 授予的高性能计算

资源和支持。

References

[1] Nikita Araslanov and Stefan Roth. Self-supervised augmentation consistency for adapting semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15384–15394, 2021.

[2] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018.

[3] Minghao Chen, Hongyang Xue, and Deng Cai. Domain adaptation for semantic segmentation with maximum squares loss. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2090–2099, 2019.

[4] Mu Chen, Zhedong Zheng, Yi Yang, and Tat-Seng Chua. Pipa: Pixel-and patch-wise self-supervised learning for domain adaptative semantic segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1905–1914, 2023.

[5] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018.

[6] Yuhua Chen, Haoran Wang, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Scale-aware domain adaptive faster r-cnn. *International Journal of Computer Vision*, 129(7):2223–2243, 2021.

[7] Seokju Cho, Heeseong Shin, Sunghwan Hong, Seungjun An, Seungjun Lee, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. *arXiv preprint arXiv:2303.11797*, 2023.

[8] Seun-An Choe, Ah-Hyung Shin, Keon-Hee Park, Jinwoo Choi, and Gyeong-Moon Park. Open-set domain adaptation for semantic segmentation. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 23943–23953, 2024.

[9] Sanjoy Chowdhury, Sayan Nag, and Dinesh Manocha. Apollo: Unified adapter and prompt learning for vision language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10173–10187, 2023.

[10] MMSegmentation Contributors. Mmsegmentation: Openmmlab semantic segmentation toolbox and benchmark, 2020.

[11] Yulu Gan, Yan Bai, Yihang Lou, Xianzheng Ma, Renrui Zhang, Nian Shi, and Lin Luo. Decorate the newcomers: Visual domain prompt for continual test time adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7595–7603, 2023.

[12] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[13] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024.

[14] Yunhe Gao, Xingjian Shi, Yi Zhu, Hao Wang, Zhiqiang Tang, Xiong Zhou, Mu Li, and Dimitris N Metaxas. Visual prompt tuning for test-time domain adaptation. *arXiv preprint arXiv:2210.04831*, 2022.

[15] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

[16] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016.

[17] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain

Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

[18] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9924–9935, 2022.

[19] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In *European Conference on Computer Vision (ECCV)*, 2022.

[20] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. Mic: Masked image consistency for context-enhanced domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11721–11732, 2023.

[21] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[22] Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. Lorahub: Efficient cross-task generalization via dynamic lora composition. *arXiv preprint arXiv:2307.13269*, 2023.

[23] Muhammad Gul Zain Ali Khan, Muhammad Ferjad Naeem, Luc Van Gool, Didier Stricker, Federico Tombari, and Muhammad Zeshan Afzal. Introducing language guidance in prompt-based continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11463–11473, 2023.

[24] Jogendra Nath Kundu, Akshay Kulkarni, Amit Singh, Varun Jampani, and R Venkatesh Babu. Generalize then adapt: Source-free domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7046–7056, 2021.

[25] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

[26] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022.

[27] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

[28] Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. Cross-domain adaptive teacher for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7581–7590, 2022.

[29] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023.

[30] Yuang Liu, Wei Zhang, and Jun Wang. Source-free domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1215–1224, 2021.

[31] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015.

[32] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31, 2018.

[33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[34] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 891–898, 2014.

[35] Muhammad Ferjad Naeem, Yongqin Xian, Xiaohua Zhai, Lukas Hoyer, Luc Van Gool, and Federico Tombari. Silc: Improving vision language pretraining with self-distillation. In *European Conference on Computer Vision*, pages 38–55. Springer, 2025.

[36] Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. Adapterfusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2005.00247*, 2020.

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[38] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6148–6157, 2017.

[39] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6956–6965, 2019.

[40] Jameel Hassan Abdul Samadh, Hanan Gani, Noor Hazim Hussein, Muhammad Uzair Khattak, Muzammal Naseer, Fahad Khan, and Salman Khan. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[41] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 443–450. Springer, 2016.

[42] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

[43] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3835–3844, 2022.

[44] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

[45] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

[46] Qin Wang, Dengxin Dai, Lukas Hoyer, Luc Van Gool, and Olga Fink. Domain adaptive semantic segmentation with self-supervised depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8515–8525, 2021.

[47] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation, 2022.

[48] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. *arXiv preprint arXiv:2310.01403*, 2023.

[49] Bin Xie, Jiale Cao, Jin Xie, Fahad Shahbaz Khan, and Yanwei Pang. Sed: A simple encoder-decoder for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[50] Binhui Xie, Shuang Li, Mingjia Li, Chi Harold Liu, Gao Huang, and Guoren Wang. Sepico: Semantic-guided pixel contrast for domain adaptive semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):9004–9021, 2023.

[51] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2945–2954, 2023.

[52] Linyan Yang, Lukas Hoyer, Mark Weber, Tobias Fischer, Dengxin Dai, Laura Leal-Taixé, Marc Pollefeys, Daniel Cremers, and Luc Van Gool. Micdrop: masking image and depth features via complementary dropout for domain-adaptive semantic segmentation. In *European Conference on Computer Vision*, pages 329–346. Springer, 2025.

[53] Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. In *The Eleventh International Conference on Learning Representations*, 2023.

[54] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019.

[55] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022.

[56] Jinghao Zhou, Li Dong, Zhe Gan, Lijuan Wang, and Furu Wei. Non-contrastive learning meets language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11028–11038, 2023.

[57] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022.

[58] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[59] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11175–11185, 2023.

[60] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018.

Lost in Translation? Vocabulary Alignment for Source-Free Domain Adaptation in Open-Vocabulary Semantic Segmentation – Supplementary Material

Silvio Mazzucco[1,2*]
silvio.mazzucco@thegoodailab.org
Carl Persson[1*]
carl.persson@thegoodailab.org
Mattia Segu[1,2,3]
mattia.segu@thegoodailab.org
Pier Luigi Dovesi[1]
pier@thegoodailab.org
Federico Tombari[3,4]
tombari@google.com
Luc Van Gool[5]
vangool@vision.ee.ethz.ch
Matteo Poggi[1,6]
m.poggi@unibo.it

[1] The Good AI Lab
thegoodailab.org
[2] ETH Zurich
[3] Google
[4] Technical University of Munich
[5] INSAIT, Sofia University, St. Kliment Ohridski
[6] University of Bologna

* joint first authorship

Introduction

This document provides supplementary material for BMVC paper "Lost in Translation? Vocabulary Alignment for Source-Free Adaptation in Open-Vocabulary Semantic Segmentation." The supplementary content elaborates on key aspects of the work, addressing implementation details, detailed analyses, and additional results.

Specifically, this document includes:

- **ChatGPT Prompts and Concept Augmentation:** Insights into the prompt engineering process and examples of how generated concepts enhance model performance.
- **Analysis of Ego-Vehicle Faulty Segmentation:** Examination of common segmentation errors, particularly for ego-vehicle-related regions, and the applied solutions.
- **Class-by-Class Analysis:** Detailed per-class performance evaluation across multiple datasets, highlighting areas of significant improvement.
- **Qualitative Results:** Visual comparisons illustrating the model's improvements in segmentation quality on various datasets.

ChatGPT o1 Prompt

Context: An open vocabulary semantic segmentation model was trained on a dataset with a set amount of labels. The model is based on CLIP, which makes it possible to extend this model to other datasets with other labels. However, the labels of the new evaluation dataset might not properly align with the labels used for training. The following task aims to solve this problem through the modification of the labels used for evaluation.

Task: Given a list of training labels and a list of evaluation labels, the objective is to find a good mapping of training labels to the evaluation labels. Multiple training labels can be mapped to the same evaluation label, but one training label can not appear more than once.

If the evaluation label does not have an exact match, then map two more synonyms, chosen freely, beyond the already mapped training labels. Also modify the evaluation label to align with the description if necessary.

Evaluation labels (with descriptions):
...

Training labels (with descriptions):
...

Follow-up ChatGPT o1 Prompt

For all evaluation labels, find a synonym and add it to the list. The synonym does not have to be from any previous list and can be chosen completely freely.

Figure 1: **Initial and Follow-up prompts used in ChatGPT o1.**

1 ChatGPT Prompts and Concept Augmentation

In this section, we describe how we obtain augmented concepts to enhance teacher performance during adaptation.

1.1 Prompt Design

To automatize the generation of augmented concepts, we interact with ChatGPT o1 making it responsible of the concepts generation process. To generate favourable concepts we feed ChatGPT with the following information: the context of the task with its corresponding restrictions, as well as information on the target labels and the source training labels. The task consists of finding a good mapping from training labels to the target labels. If this mapping does not exist, ChatGPT is asked to generate synonyms instead.

We also found that the descriptions of the labels in the evaluation datasets did not align well with the visual semantics of such labels. In this case, we also ask ChatGPT to find a suitable replacement label that aligns better with the label description. In practice, this was seldom the case but helped in certain edge case scenarios. We found that this prompt could generate satisfactory concepts for all datasets. However, for larger datasets, the amounts of synonyms to generate was changed to satisfy computational restrictions, but only slight

Original Classes	ChatGPT Concepts
road	street
sidewalk	pavement, floor-stone, floor-tile, footpath
building	building-other, house, skyscraper, structure
wall	wall-other, solid-other, structural-other, barrier
fence	railing, enclosure
pole	light, metal, post
traffic light	signal
traffic sign	stop sign, parking meter, banner, signboard
vegetation	tree, bush, plant-other, foliage
terrain	grass, ground-other, dirt, landscape
sky	sky-other, heavens
person	human
rider	horse, skateboard, cyclist
car	automobile
truck	lorry
bus	coach
train	locomotive
motorcycle	motorbike
bicycle	bike

Table 1: **ChatGPT Concepts.** List of the original classes and the additional concepts generated by ChatGPT o1 for CityScapes.

Method	road	swalk	build.	wall	fence	pole	tlight	sign	veg.	terrain	sky	person	rider	car	truck	bus	train	mbike	bike	mIoU
Zero-Shot	**86.03**	48.58	81.79	9.77	**38.37**	27.47	46.09	**50.05**	84.08	0.02	81.69	68.08	**0.05**	**76.38**	23.22	49.18	7.12	**55.35**	**70.23**	47.56
Ours	84.87	**50.70**	**84.90**	**34.32**	31.70	**37.71**	**47.59**	49.65	**87.12**	**45.95**	**88.70**	**69.06**	0.00	67.48	**46.76**	**52.35**	**16.28**	54.83	69.72	**53.67**
	-1.16	+2.12	+3.11	+24.55	-6.67	+10.24	+1.50	-0.40	+3.04	+45.93	+7.01	+0.98	-0.05	-8.9	+23.54	+3.17	+9.16	-0.52	-0.51	+6.11

Table 2: **Per-class results on CityScapes.** We compare our proposed method to the initial Zero-Shot predictions by CAT-Seg.

variations to the prompt were made.

1.2 Examples of Augmented Concepts

By exploiting the prompts introduced before, we can generate augmented concepts related to any of the semantic classes in a dataset. In Table 1, we show some examples of concepts generated by ChatGPT o1 specifically for the original semantic classes in the CityScapes dataset.

2 Ego-Vehicle Faulty Segmentation

For the sake of completeness, in Table 2 we report the results concerning all of the single classes on CityScapes. As mentioned in the main paper, we observe a weird worsening of the performance on the class *car* in the CityScapes results. We assume that this mainly

Method	road	swalk	build.	wall	fence	pole	tlight	sign	veg.	terrain	sky	person	rider	car	truck	bus	train	mbike	bike	mIoU
Zero-Shot Crop	**89.2**	55.25	82.3	13.52	**38.55**	27.51	46.09	51.26	84.45	0.02	86.58	68.12	**0.05**	81.19	24.13	49.65	13.92	**55.48**	**70.99**	49.38
Ours Crop	88.91	**57.06**	**85.97**	**35.89**	31.65	**38.77**	**48.69**	**54.56**	**87.2**	**47.07**	**90.14**	**69.11**	0.0	**85.67**	**47.1**	**53.31**	**17.34**	55.17	70.15	55.99
	-0.29	+1.81	+3.67	+22.37	-6.90	+11.26	+2.60	+3.30	+2.75	+47.05	+3.56	+0.99	-0.05	+4.48	+22.97	+3.66	+3.42	-0.31	-0.84	+6.61

Table 3: **Per-class results on CityScapes when cropping.** We compare our proposed method to the initial Zero-Shot predictions by CAT-Seg when cropping the images to get rid of the ego-vehicle during evaluation.

Method	plane	bag	bed	bedcloth.	bench	bike	bird	boat	book	bottle	build	bus	cabinet	car	cat	ceil.	chair	cloth	computer	mIoU
Zero-Shot	88.82	**43.08**	**25.47**	40.42	26.78	76.64	90.35	76.47	52.54	79.49	**42.19**	94.24	42.32	74.69	91.21	**52.76**	**56.35**	14.00	14.96	55.52
Ours	**89.02**	41.78	25.43	**42.61**	**27.42**	**77.04**	**90.68**	**78.66**	**56.89**	**79.68**	39.18	**94.37**	**43.18**	**79.08**	**91.90**	50.50	47.07	**19.39**	**18.25**	57.01
	+0.20	-1.30	-0.04	+2.19	+0.64	+0.4	+0.33	+2.19	+4.35	+0.19	-3.01	+0.13	+0.86	+4.39	+0.69	-2.26	-9.28	+5.39	+3.29	+1.49

Method	cow	cup	curtain	dog	door	fence	floor	flower	food	grass	ground	horse	kboard	light	mbike	mountain	mouse	person	plate	pform	mIoU
Zero-Shot	90.19	**36.79**	**56.02**	89.01	25.00	40.63	62.16	**33.63**	39.15	79.8	22.86	89.78	**82.31**	46.2	87.28	58.64	45.3	87.31	6.74	32.95	55.52
Ours	**90.61**	36.31	55.54	**89.67**	**25.01**	**40.84**	**66.12**	21.90	**44.04**	**81.82**	**40.35**	**90.37**	81.14	**46.93**	**88.01**	**59.39**	**50.15**	**87.37**	**8.45**	**40.63**	57.01
	+0.42	-0.48	-0.48	+0.66	+0.01	+0.21	+3.96	-11.73	+4.89	+2.02	+17.49	+0.59	-1.17	+0.73	+0.73	+0.75	+4.85	+0.06	+1.71	+7.68	+1.49

Method	plant	road	rock	sheep	shelves	s.walk	sign	sky	snow	sofa	table	track	train	tree	truck	tv	wall	water	window	wood	mIoU
Zero-Shot	59.91	49.07	47.22	90.34	28.69	13.55	29.21	93.76	**74.71**	**45.84**	**57.81**	33.74	83.43	**78.89**	11.33	74.55	**62.84**	89.36	**39.55**	17.5	55.52
Ours	**61.09**	**50.92**	**51.61**	**90.57**	**32.13**	**19.28**	**31.77**	**93.97**	74.11	40.04	56.86	**52.04**	**87.90**	78.46	**13.91**	**77.34**	61.91	**90.72**	37.91	**24.14**	57.01
	+1.18	+1.85	+4.39	+0.23	+3.44	+5.73	+2.56	+0.21	-0.6	-5.8	-0.95	+18.3	+4.47	-0.43	+2.58	+2.79	-0.93	+1.36	-1.64	+6.64	+1.49

Table 4: **Per-class results on PASCAL-Context 59.** We compare our proposed method to the Zero-Shot predictions by CAT-Seg.

Method	chand.	bus	barrel	bridge	poster	schair	wpane	tank	palm	howel	k.island	tower	fountain	ball	boat	build.	rug	shelf	pot	skyscraper	mIoU
Zero-Shot	33.83	71.00	6.36	34.93	7.79	4.34	42.19	1.43	25.81	0.14	**8.29**	**21.60**	**36.23**	**54.75**	**65.93**	**60.58**	**46.47**	**27.10**	**24.39**	**25.75**	26.88
Ours	**51.37**	**85.17**	**15.75**	**43.32**	**14.76**	**10.85**	**48.41**	**7.60**	**31.15**	**5.47**	3.53	16.14	29.90	47.41	57.94	52.24	38.06	17.17	13.82	14.91	27.39
	+17.54	+14.17	+9.39	+8.39	+6.97	+6.51	+6.22	+6.17	+5.34	+5.33	-4.76	-5.46	-6.33	-7.34	-7.99	-8.34	-8.41	-9.93	-10.57	-10.84	+0,51

Table 5: **Top 10 best and worst per-class results on ADE20k** We compare our proposed method to the Zero-Shot predictions by CAT-Seg.

comes from the ego-vehicle, which is present during both training and evaluation. We thus decide to run an evaluation in which we crop out the bottom 23% of the image, ensuring the ego-vehicle is completely removed from every image. We run this evaluation both before and after the adaptation, and collect results in Table 3. We can notice that the ego-vehicle worsens the performance for the class *car* also for the original Cat-SEG model – since the mIoU already rises from 76.38 (see main paper) to 81.19. Furthermore, by running our adaptation strategy in this setting, our result on this class gets even better, reaching a score of 85.67.

3 Class-by-Class Analysis

In the main paper, on PASCAL-Context 59 and ADE20k we only reported the mIoU averaged over all semantic classes for the sake of space. Here, we report more detailed results concerning single classes.

Table 4 collects the mIoU for each single semantic class in PASCAL-Context 59. We can appreciate how the majority of the classes are improved by our method, with only 15 out of the total 59 classes showing some drops.

Table 5 shows the results concerning the 20 classes for which we had the highest increase/drop in mIoU on ADE20k. We can appreciate how classes such as *bus* and *chand.* are largely improved by more than 10%. As a counterpoint, others such as *pot* and *skyscrapers* show a similar drop in performance.

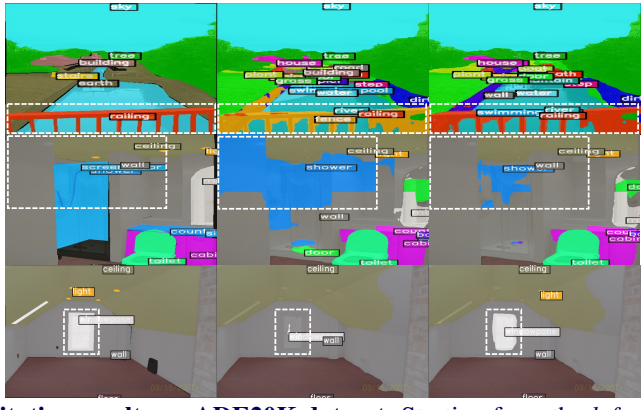

Figure 2: **Qualitative results on ADE20K dataset.** Starting from the *left*, we show respectively the ground truth segmentation map, the zero-shot CAT-Seg prediction, and the result after our adaptation process.

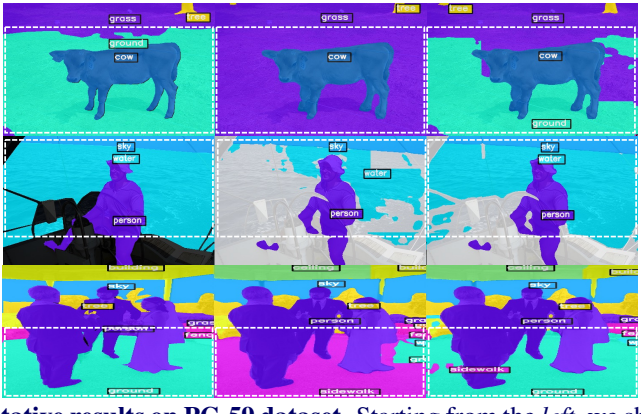

Figure 3: **Qualitative results on PC-59 dataset.** Starting from the *left*, we show respectively the ground truth segmentation map, the zero-shot CAT-Seg prediction, and the result after our adaptation process.

4 Qualitative Results

We conclude with some additional qualitative results.

Figure 2 shows some examples from the ADE20k dataset over three rows, respectively overlayed, from left to right, with ground-truth semantic masks, semantic labels predicted by the original CAT-Seg model, or by the one adapted with our strategy. In the first row, we can appreciate how the adaptation process allows CAT-Seg to fully recover the railing semantic class, which was almost lost by the zero-shot model. In the second example, the original CAT-Seg model wrongly labelled a large portion of the scene as part of the *shower* class, whereas it can limit such area to the actual shower visible in the image. Finally, in the third example, we can appreciate how CAT-Seg can properly label even very small objects in the scene – such as the *light* class appearing on the roof, which was completely ignored by the original CAT-Seg weights.

Figure 2 reports three examples from PASCAL-Context 59. Here, in particular, we can appreciate how the adaptation process allows for recovering large regions of the background, being incorrectly labeled as *grass* (top) or *sidewalk* (bottom) by the zero-shot model.