Lost in Translation? Vocabulary Alignment for Source-Free Adaptation in Open-Vocabulary Semantic Segmentation

Silvio Mazzucco[1] Carl Persson[2] Mattia Segu[1] Pier Luigi Dovesi[2]
Federico Tombari[1] Luc Van Gool[2] Matteo Poggi[1]

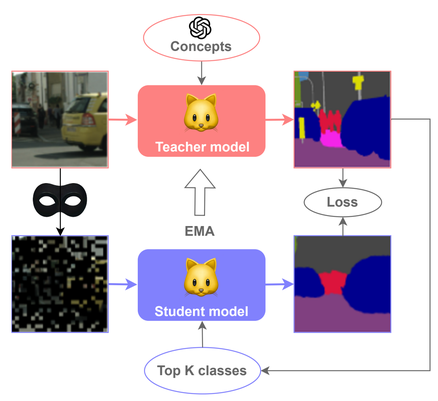
Project page: `https://thegoodailab.org/blog/vocalign`

BMVC 2025

Introduction

Challenges. Open-Vocabulary Semantic Segmentation models may suffer from domain-shift issues both in terms of visual content and vocabulary misalignments. Adaptation approaches requiring access to their training data are impractical.

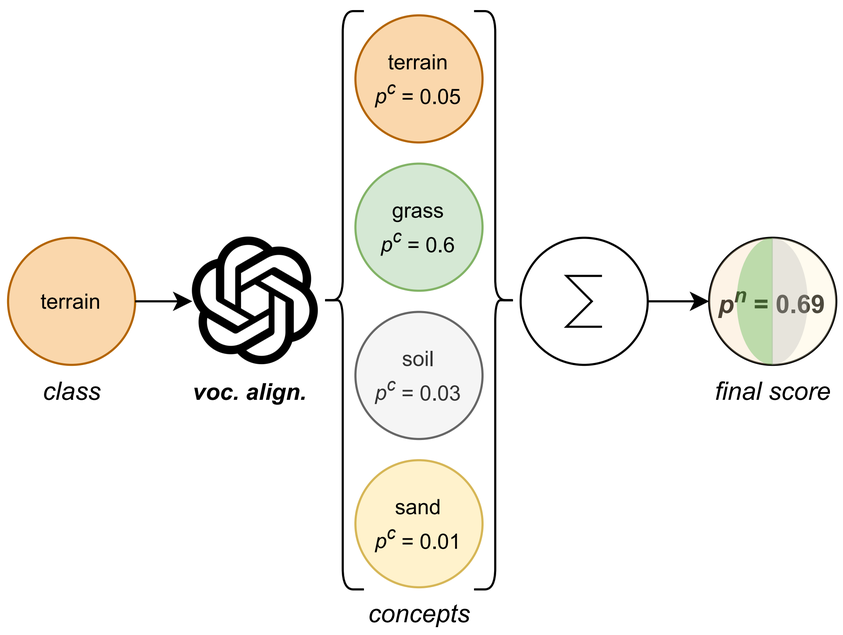
Proposal. `VocAlign`, a source-free domain adaptation framework handling both visual and vocabulary -related domain shifts.

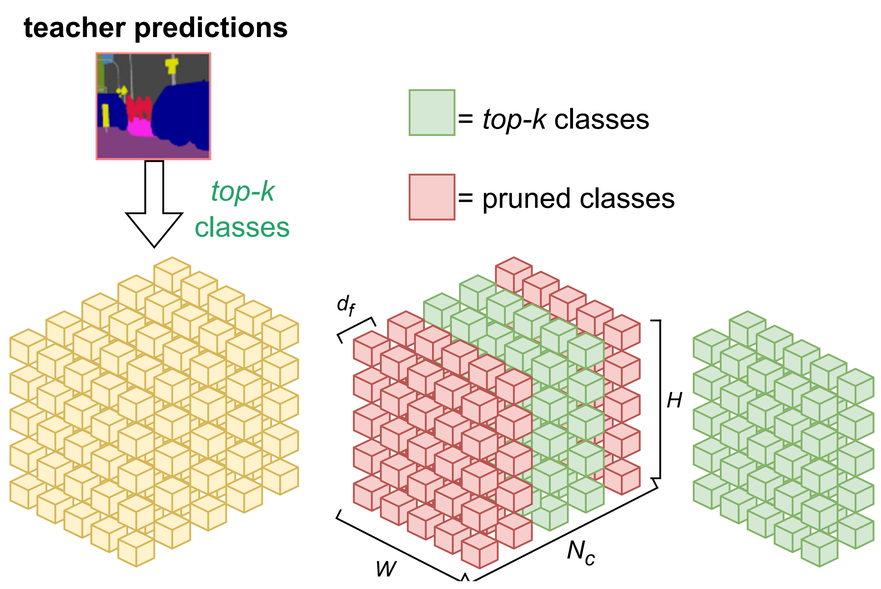


Methodology

Teacher-Student framework. We apply VocAlign on top of the CAT-Seg backbone. We implement a teacher-student framework, where the teacher is updated over time as an exponential moving average (EMA) of the student weights, and the student processes masked images.

On top of it, we introduce two key components specifically designed to deal with the Open-Vocabulary setting: **Vocabulary Alignment** and **Top-K Selection**.





• **Vocabulary Alignment.** We enrich each class in the target dataset with $c(n)$ *concepts* – additional text descriptions or synonyms, used only during adaptation. Before pseudo-label generation, concepts are aggregated back:

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

• **Top-K Selection.** Adapting to many classes can be computationally prohibitive. To address this, we use teacher predictions (enhanced by concepts) to identify likely classes in an image. Given teacher probabilities $\hat{y} \in \mathbb{R}^{N_c,H,W}$, we compute the mean activation per class and select the *Top-K* with the highest averages. The remaining classes are pruned from the cost volume before aggregation.

Quantitative Results

CityScapes evaluation

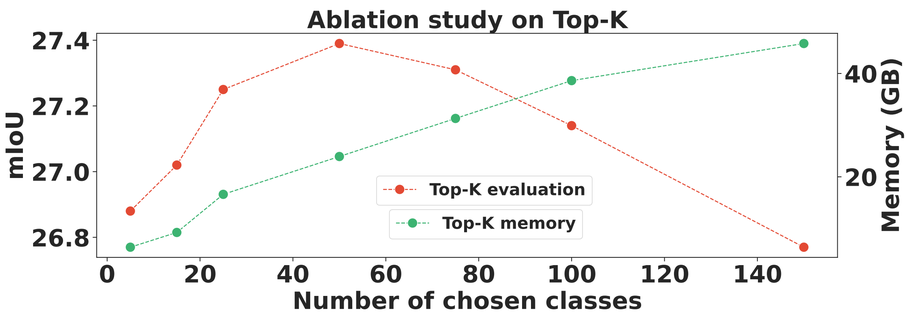
Method	car	fence	road	mbike	...	pole	truck	wall	terrain	mIoU
Zero-Shot CAT-Seg	**76.38**	**38.37**	**86.03**	**55.35**	...	27.47	23.22	9.77	0.02	47.56
VocAlign (ours)	67.48	31.70	84.87	54.83	...	37.71	**46.76**	**34.32**	**45.95**	**53.67**
Improve (%)	-8.9	-6.67	-1.16	-0.52	...	+10.24	+23.54	+24.55	+45.93	+6.11

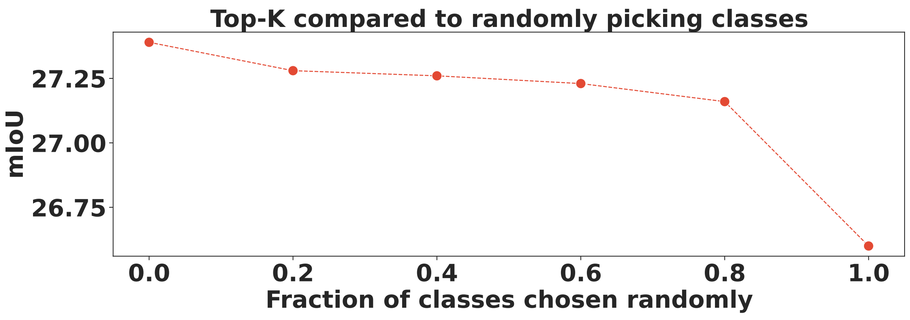
Multi-dataset evaluation

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**
Improve (%)	+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**
Improve (%)	+1.09	+0.38	+0.18

Ablation studies





Qualitative Results

CityScapes dataset

PC-59 dataset