Boosting Unsupervised Semantic Segmentation with Principal Mask Proposals

Oliver Hahn [1] Nikita Araslanov [2,3] Simone Schaub-Meyer [1,4] Stefan Roth [1,4]

[1] TU Darmstadt [2] TU Munich [3] Munich Center for Machine Learning [4] hessian.AI

CVPR JUNE 17-21, 2024 SEATTLE, WA

Paper

Introduction

Task: Unsupervised semantic segmentation (USS) aims to consistently discover and categorize image regions in a given data domain without any labels.

Motivation:

- Can we directly use the potential of pre-trained self-supervised features instead of learning a new representation on top?
- Large performance gap comparing supervised and unsupervised probing within pre-trained self-supervised feature spaces suggests hidden potential.

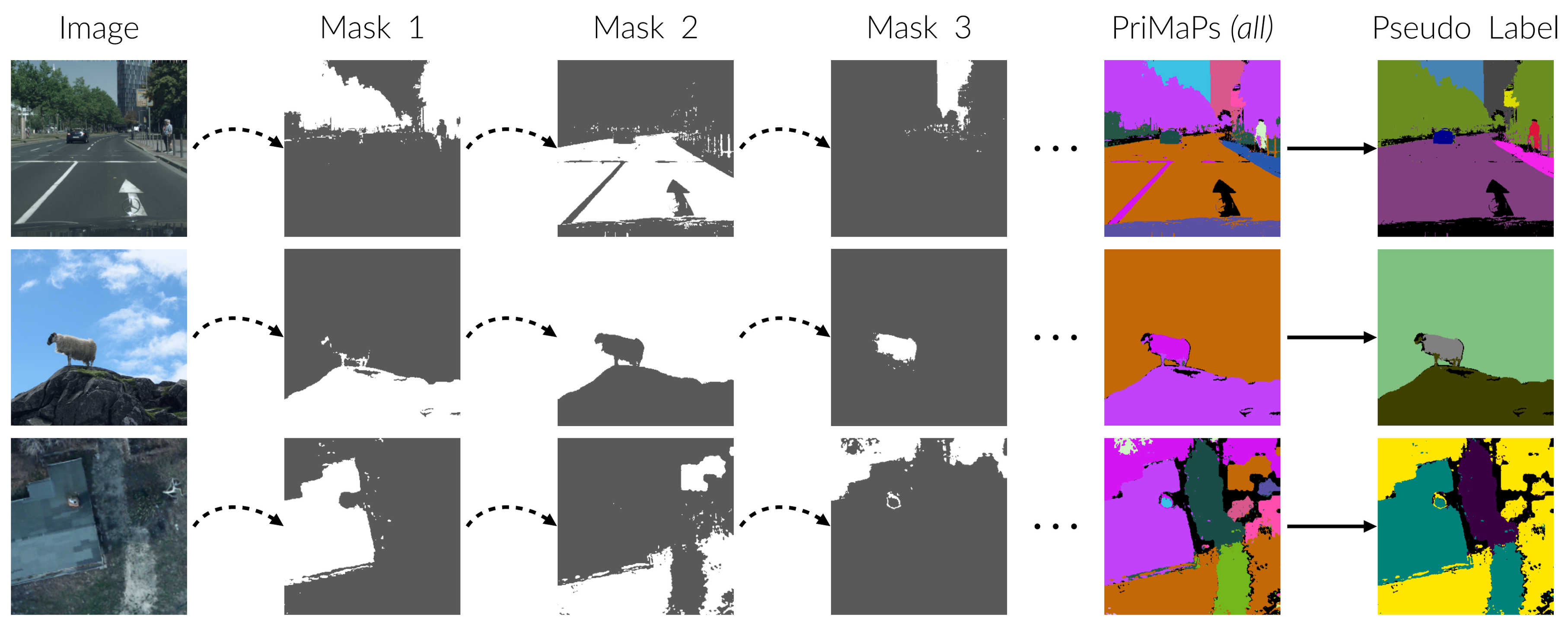


Figure 1. **PriMaPs divide images into masks. Assigning a pseudo ID per mask leads to pseudo labels.**

Idea: Use principal components of self-supervised features to identify visual patterns with high semantic correlation to decompose images into mask proposals. Construct pseudo labels and directly optimize class prototypes for USS in the feature space.

References & Acknowledgments

[1] Mathilde Caron et al. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.
[2] Mark Hamilton et al. Unsupervised semantic segmentation by distilling feature correspondences. In *ICLR*, 2022.
[3] Kehan Li et al. Dynamic clustering network for unsupervised semantic segmentation. In *CVPR*, 2023.
[4] Maxime Oquab et al. DINOv2: Learning robust visual features without supervision. In *TMLR*, 2024.
[5] Hyun Seok Seong et al. Leveraging hidden positives for unsupervised semantic segmentation. In *CVPR*, 2023.
[6] Zhaoyuan Yin et al. TransFGU: A top-down approach to fine-grained unsupervised semantic segmentation. In *ECCV*, 2022.

This project is partially funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 866008) as well as the State of Hesse (Germany) through the cluster projects "The Third Wave of Artificial Intelligence (3AI)" and "The Adaptive Mind (TAM)".

TL;DR

We present PriMaPs – Principal Mask Proposals – decomposing images into semantically meaningful masks based on their feature representation. This enables unsupervised semantic segmentation by fitting class prototypes to PriMaPs with stochastic expectation-maximization.

Method

PriMaPs iteratively decompose images into class-agnostic mask proposals based on self-supervised representations.

With dense features $f \in \mathbb{R}^{C\times H\times W}$ for every mask proposal P:

1. Nearest neighbor feature $\tilde{f}$ of first principal component v_1
2. Cosine-distance similarity map:
$$M = (M_{i,j})_{i,j}\,,\text{where } M_{i,j} = \big(\tilde{f}\big)^{\top}\hat{f}_{:,i,j}$$
3. Principal mask with $\psi \in (0,1)$:
$$P = \Big[M_{i,j} > \psi \cdot \max_{m,n} M_{m,n}\Big]_{i,j}$$

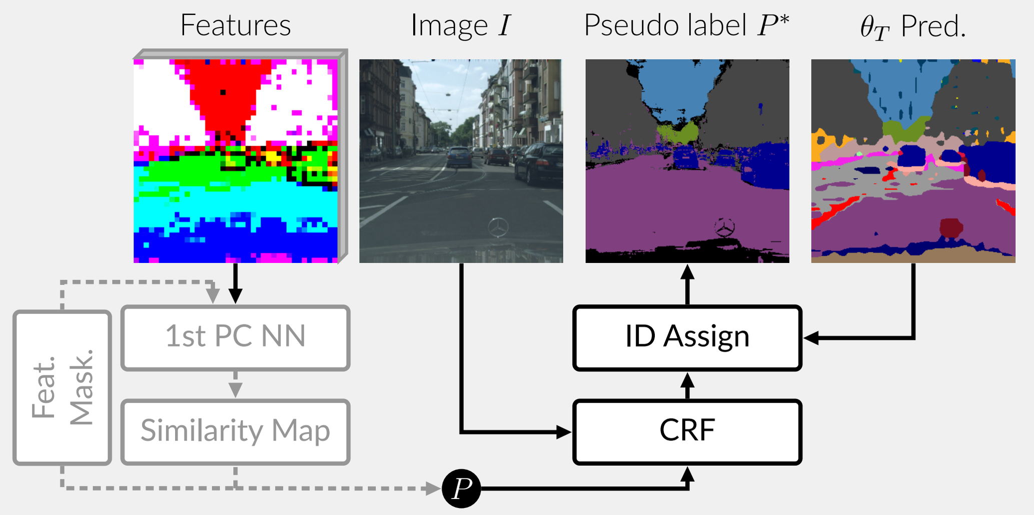


Figure 2. **PriMaPs pseudo label generation.**

PriMaPs-EM fits class prototypes by optimizing over two identically sized vector sets using stochastic EM of a clustering objective guided by PriMaPs.

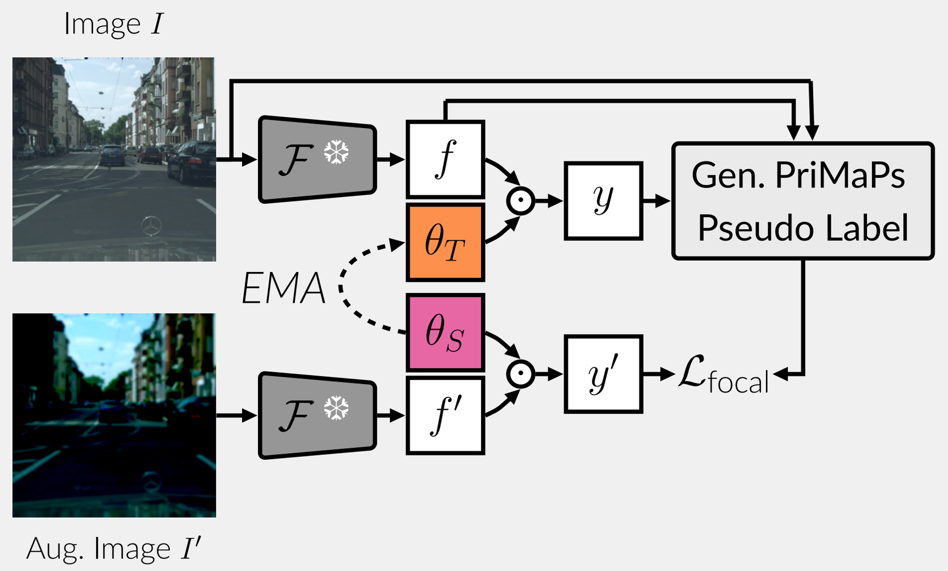


Figure 3. **PriMaPs-EM architecture.**

Initialize class prototypes θ with cosine-distance batch-wise K-means loss:

$$\mathcal{L}_{K\text{-means}}(\theta_T) = -\sum_{i,j}\max\big(\theta_T^{\top} f_{:,i,j}\big)$$

Further optimize with focal loss:

$$\mathcal{L}_{\text{focal}}(\theta_S; y') = -\sum_{k,i,j}(1-\chi_k)^2 P^*_{k,i,j}\log(y'_{k,i,j})$$

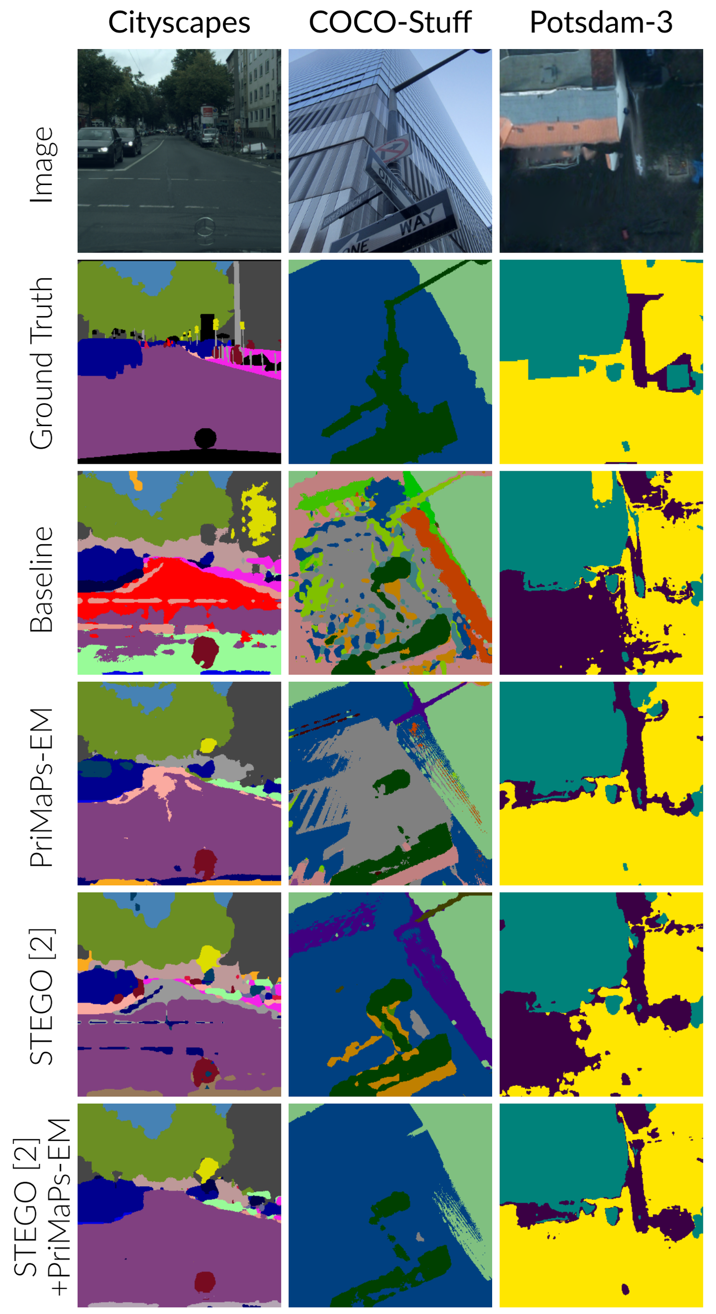
with $y'_{:,i,j} = \mathrm{softmax}(\theta_S^{\top} f'_{:,i,j})\,,$
and class-wise confidence χ_k

Results

Table 1. **Comparison to existing unsupervised semantic segmentation methods**, using Accuracy and mean IoU (in %).

Method	Backbone	*Cityscapes*		*COCO-Stuff*		*Potsdam-3*	
		Acc	mIoU	Acc	mIoU	Acc	mIoU
Baseline [1]	DINO ViT-S/8	61.4	15.8	34.2	9.5	56.6	33.6
+ TransFGU [6]		77.9	16.8	52.7	17.5	–	–
+ STEGO [2]		–	–	48.3	24.5	77.0	62.6
+ ACSeg [3]		–	–	–	16.4	–	–
+ HP [5]		80.1	18.4	57.2	24.6	–	–
+ PriMaPs-EM		**81.2**	**19.3**	46.5	16.4	62.5	39.0
+ SotA + PriMaPs-EM		76.3	19.2	**57.8**	**25.1**	**78.4**	**64.2**
Baseline [1]	DINO ViT-B/8	49.2	15.5	38.8	15.7	66.1	49.4
+ STEGO [2]		73.2	21.0	56.9	28.2	–	–
+ HP [5]		**79.5**	18.4	–	–	82.4	69.1
+ PriMaPs-EM		59.6	17.6	48.4	21.9	80.5	66.9
+ SotA + PriMaPs-EM		78.6	**21.6**	**57.9**	**29.7**	**83.3**	**71.0**
Baseline [4]	DINOv2	49.5	15.3	44.5	22.9	75.9	61.0
+ PriMaPs-EM	ViT-S/14	**71.6**	**19.0**	**46.4**	**23.8**	**78.4**	**64.2**
Baseline [4]	DINOv2	36.1	14.9	35.0	17.9	82.4	69.9
+ PriMaPs-EM	ViT-B/14	**82.8**	**21.2**	**52.6**	**23.6**	**83.1**	**71.0**

Figure 4. **Qualitative results** for DINO ViT-B/8.



Experiments: PriMaPs-EM provides modest but consistent improvements across all settings. Qualitative results indicate improved local segmentation consistency.

Summary

- **Lightweight mask proposals**, leveraging intrinsic properties of the embedding space provided by an off-the-shelf self-supervised learning approach.
- **Pseudo labels** based on the mask proposals, and a straightforward stochastic expectation-maximization **approach for boosting USS**.
- **Improved USS results** across a wide range of self-supervised embeddings and datasets as well as orthogonal to current SotA methods.