Boosting Unsupervised Semantic Segmentation with Principal Mask Proposals

Anonymous authors
Paper under double-blind review

Abstract

Unsupervised semantic segmentation aims to automatically partition images into semantically meaningful regions by identifying global categories within an image corpus without any form of annotation. Building upon recent advances in self-supervised representation learning, we focus on how to leverage these large pre-trained models for the downstream task of unsupervised segmentation. We present PriMaPs – Principal Mask Proposals – decomposing images into semantically meaningful masks based on their feature representation. This allows us to realize unsupervised semantic segmentation by fitting class prototypes to PriMaPs with a stochastic expectation-maximization algorithm, PriMaPs-EM. Despite its conceptual simplicity, PriMaPs-EM leads to competitive results across various pre-trained backbone models, including DINO and DINOv2, and across datasets, such as Cityscapes, COCO-Stuff, and Potsdam-3. Importantly, PriMaPs-EM is able to boost results when applied orthogonally to current state-of-the-art unsupervised semantic segmentation pipelines.

1 Introduction

Semantic image segmentation is a dense prediction task that classifies image pixels into categories from a pre-defined semantic taxonomy. Owing to its fundamental nature, semantic segmentation has a broad range of applications, such as image editing, medical imaging, robotics, or autonomous driving (see Minaee et al., 2022, for an overview). Addressing this problem via supervised learning requires ground-truth labels for every pixel (Long et al., 2015; Ronneberger et al., 2015; Chen et al., 2018b). Such manual annotation is extremely time and resource intensive. For instance, a trained human annotator requires an average of 90 minutes to label up to 30 classes in a single 2 MP image (Cordts et al., 2016). While committing significant resources to large-scale annotation efforts achieves excellent results (Kirillov et al., 2023), there is natural interest in a more economical approach. Alternative lines of research aim to solve the problem using cheaper – so-called "weaker" – variants of annotation. For example, image-level supervision describing the semantic categories present in the image, or bounding-box annotations, can reach impressive levels of segmentation accuracy (Dai et al., 2015; Araslanov & Roth, 2020; Oh et al., 2021; Xu et al., 2022; Ru et al., 2023).

As an extreme problem scenario toward reducing the annotation effort, unsupervised semantic segmentation aims to consistently discover and categorize image regions in a given data domain without any labels, knowing only how many classes to discover. Unsupervised semantic segmentation is highly ambiguous as class boundaries and the level of categorical granularity are task-dependent.[1] However, we can leverage the fact that typical image datasets have a homogeneous underlying taxonomy and exhibit invariant domain characteristics. Therefore, it is still feasible to decompose images in such datasets in a semantically meaningful and consistent manner without annotations.

Despite the challenges of unsupervised semantic segmentation, we have witnessed remarkable progress on this task in the past years (Ji et al., 2019; Cho et al., 2021; Van Gansbeke et al., 2021; 2022; Ke et al., 2022; Yin et al., 2022; Hamilton et al., 2022; Karlsson et al., 2022; Li et al., 2023; Seong et al., 2023; Seitzer et al., 2023). Deep representations obtained with self-supervised learning (SSL), such as DINO (Caron et al., 2021),

[1]While assigning actual semantic labels to regions without annotation is generally infeasible, the assumption is that the categories of the discovered segments will strongly correlate with human notions of semantic meaning.

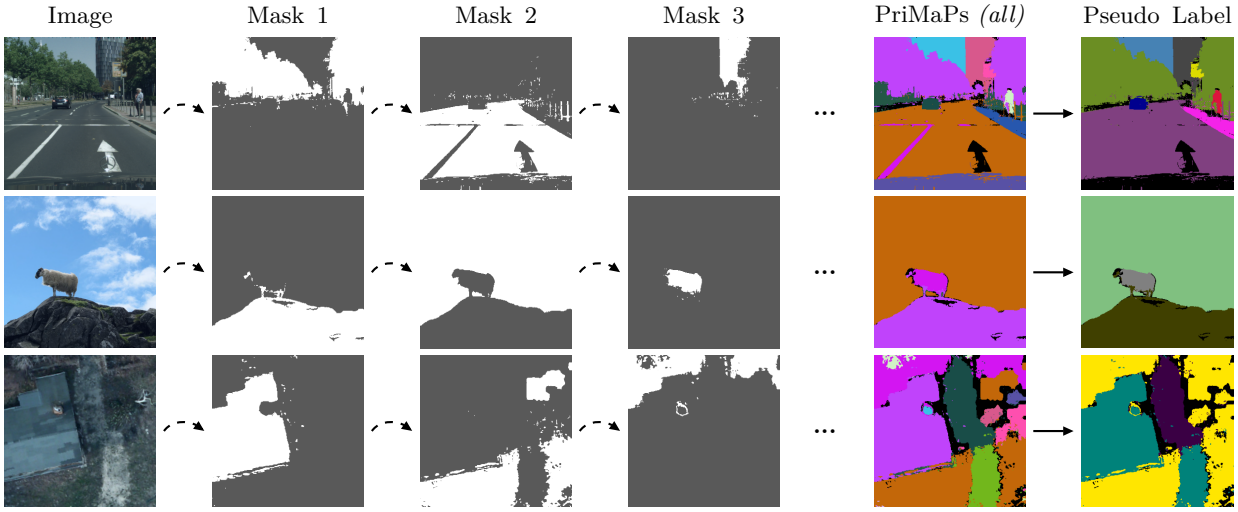


Figure 1: **PriMaPs pseudo label example.** Principal mask proposals (PriMaPs) are iteratively extracted from an image (dashed arrows). Each mask is assigned a semantic class resulting in a pseudo label. The examples are taken from the Cityscapes (top), COCO-Stuff (middle), and Potsdam-3 (bottom) datasets.

have played a critical role in this advance. However, it remains unclear whether previous work leverages the intrinsic properties of the original SSL representations, or merely uses them for "bootstrapping" and learns a new representation on top. Exploiting the inherent properties of SSL features is preferable for two reasons. First, training SSL models incurs a substantial computational effort, justifiable only if the learned feature extractor is sufficiently versatile. In other words, one can amortize the high computational cost over many downstream tasks, provided that task specialization is computationally negligible. Second, studying SSL representations with lightweight tools, such as linear models, leads to a more interpretable empirical analysis than with the use of more complex models, as evidenced by the widespread use of linear probing in SSL evaluation. Such interpretability advances research on SSL models toward improved cross-task generalization.

Equipped with essential tools of linear modeling, *i. e.* Principal Component Analysis (PCA), we generate **Pri**ncipal **Ma**sk **P**roposal**s**, or PriMaPs, directly from the SSL representation. Complementing previous findings on object-centric images (Tumanyan et al., 2022; Amir et al., 2022), we show that principal components of SSL features tend to identify visual patterns with high semantic correlation also in scene-centric imagery. Leveraging PriMaPs and minimalist post-processing, we construct semantic pseudo labels for each image as illustrated in Fig. 1. Finally, instead of learning a new embedding on top of the SSL representation (Hamilton et al., 2022; Seong et al., 2023; Seitzer et al., 2023; Zadaianchuk et al., 2023), we employ a moving average implementation of stochastic Expectation Maximization (EM) (Chen et al., 2018a) to assign a consistent category to each segment in the pseudo labels and directly optimize class prototypes in the feature space. Our experiments show that this straightforward approach not only boosts the segmentation accuracy of the DINO baseline, but also that of more advanced state-of-the-art approaches tailored for semantic segmentation, such as STEGO (Hamilton et al., 2022) and HP (Seong et al., 2023).

We make the following contributions: *(i)* We derive lightweight mask proposals, leveraging intrinsic properties of the embedding space, *e. g.*, the covariance, provided by an off-the-shelf SSL approach. *(ii)* Based on the mask proposals, we construct pseudo labels and employ moving average stochastic EM to assign a consistent semantic class to each proposal. *(iii)* We demonstrate improved segmentation accuracy across a wide range of SSL embeddings and datasets.

2 Related Work

Our work builds upon recent advances in self-supervised representation learning, and takes inspiration from previous unsupervised semantic and instance segmentation methods.

The goal of **self-supervised representation learning (SSL)** is to provide generic, task-agnostic feature extractors (He et al., 2020; Chen et al., 2020; Grill et al., 2020). A pivotal role in defining the behavior of self-supervised features on future downstream tasks is taken by the self-supervised objective, the so-called *pretext* task. Examples of such tasks include predicting the context of a patch (Doersch et al., 2015) or its rotation (Gidaris et al., 2018), image inpainting (Pathak et al., 2016), and "solving" jigsaw puzzles (Noroozi & Favaro, 2016). Another family of self-supervised techniques is based on contrastive learning (Chen et al., 2020; Caron et al., 2020). More recently, Transformer networks (Dosovitskiy et al., 2020) revived some older pretext tasks, such as context prediction (Caron et al., 2021; He et al., 2022), in a more data-scalable fashion.

While the standard evaluation practice in SSL (*e. g.*, linear probing, transfer learning) offers some glimpse into the feature properties, understanding the embedding space produced by SSL remains an active terrain for research (Ericsson et al., 2021; Naseer et al., 2021). In particular, DINO features (Caron et al., 2021; Oquab et al., 2024) are known to encode accurate object-specific information, such as object parts (Amir et al., 2022; Tumanyan et al., 2022). However, it remains unclear to what extent DINO embeddings allow for semantic representation of the more ubiquitous multi-object scenes. Here, following previous work (*e. g.*, Hamilton et al., 2022; Seong et al., 2023), we provide further insights.

Early techniques for **unsupervised semantic segmentation** using deep networks (Cho et al., 2021; Van Gansbeke et al., 2021) approach the problem in the spirit of transfer learning and, under certain nomenclatures, may not be considered fully unsupervised. Specifically, starting with *supervised* ImageNet pre-training (Russakovsky et al., 2015), a network obtains a fine-tuning signal from segmentation-oriented training objectives. Such supervised "bootstrapping" appears to be crucial in the ill-posed unsupervised formulation. Unsupervised training of a deep model for segmentation from scratch is possible, albeit sacrificing accuracy (Ji et al., 2019; Ke et al., 2022). However, training a new deep model for each downstream task contradicts the spirit of SSL of amortizing the high SSL training costs over many computationally cheap specializations of the learned features (Bommasani et al., 2021).

Relying on *self-supervised* DINO pre-training, recent work (Hamilton et al., 2022; Li et al., 2023; Seong et al., 2023) has demonstrated the potential of such amortization with more lightweight fine-tuning for semantic segmentation. Nevertheless, most of this work (*e. g.*, Hamilton et al., 2022; Van Gansbeke et al., 2022) has treated the SSL representation as an inductive prior by learning a new embedding space over the SSL features (*e. g.*, Hamilton et al., 2022; Seong et al., 2023). In contrast, following SSL principles, we use the SSL representation in a more direct and lightweight fashion – by extracting mask proposals using linear models (PCA) with minimal post-processing and learning a direct mapping from feature to prediction space.

Mask proposals have an established role in computer vision (Arbelaez et al., 2011; Uijlings et al., 2013), and remain highly relevant in deep learning (Hwang et al., 2019; Van Gansbeke et al., 2021; Yin et al., 2022). Different from previous work, we directly derive the mask proposals from SSL representations. Our approach is inspired by the recent use of classical algorithms, such as normalized cuts (Ncut Shi & Malik, 2000), in the context of self-supervised segmentation (Wang et al., 2023a;b). However previous approaches (Van Gansbeke et al., 2021; 2022; Wang et al., 2023a;b) mainly proposed foreground object masks on object-centric data, utilized in a multi-step self-training. In contrast, we develop a straightforward method for extracting dense pseudo labels for learning unsupervised semantic segmentation of scene-centric data and show consistent benefits in improving the segmentation accuracy across a variety of baselines and state-of-the-art methods (Hamilton et al., 2022; Seong et al., 2023).

3 PriMaPs: Principal Mask Proposals

This work leverages recent advances in self-supervised representation learning (Caron et al., 2021; Oquab et al., 2024) for the specific downstream task of unsupervised semantic segmentation. Our approach is based on the observation that such pre-trained features already exhibit intrinsic spatial similarities, capturing semantic correlations, providing guidance to fit global pseudo-class representations.

A simple baseline. Consider a simple baseline that applies K-means clustering to DINO ViT features (Caron et al., 2021). Surprisingly, this already leads to reasonably good unsupervised semantic segmentation results, *e. g.*, around 15 % mean IoU to segment 27 classes on Cityscapes (Cordts et al., 2016), see Tab. 1.

However, supervised linear probing between the same feature space and the ground-truth labels – the theoretical upper bound – leads to clearly superior results of almost 36 %. Given this gap and the simplicity of the approach, we conclude that there is *valuable potential* in directly obtaining semantic segmentation without enhancing the original feature representation, unlike in previous work (Hamilton et al., 2022; Seong et al., 2023).

From K-means to PriMaPs-EM. When examining the K-means baseline as well as state-of-the-art methods (Hamilton et al., 2022; Seong et al., 2023), see Fig. 4, it can be qualitatively observed that more local consistency within the respective predictions would already lead to less mis-classification. We take inspiration from (Drineas et al., 2004; Ding & He, 2004), who showed that the PCA subspace, spanned by principal components, is a relaxed solution to K-means clustering. We observe that principal components have high semantic correlation for object- as well as scene-centric image features (*cf.* Fig. 1). We utilize this by iteratively partitioning images based on dominant feature patterns, identified by means of the cosine similarity of the image features to the respective first principal component. We name the resulting class-agnostic image decomposition *PriMaPs* – Principal Mask Proposals. PriMaPs stem directly from SSL representations and guide the process of unsupervised semantic segmentation. Shown in Fig. 2, our optimization-based approach, PriMaPs-EM, operates over an SSL feature representation computed from a frozen deep neural network backbone. The optimization realizes stochastic EM of a clustering objective guided by PriMaPs. Specifically, PriMaPs-EM fits class prototypes to the proposals in a globally consistent manner by optimizing over two identically sized vector sets, with one of them being an exponential moving average (EMA) of the other. We show that PriMaPs-EM enables accurate unsupervised partitioning of images into semantically meaningful regions while being highly lightweight and orthogonal to most previous approaches in unsupervised semantic segmentation.

3.1 Deriving PriMaPs

We start with a frozen pre-trained self-supervised backbone model $\mathcal{F} : \mathbb{R}^{3\times h\times w} \to \mathbb{R}^{C\times H\times W}$, which embeds an image $I \in \mathbb{R}^{3\times h\times w}$ into a dense feature representation $f \in \mathbb{R}^{C\times H\times W}$ as

$$f = \mathcal{F}(I)\,. \tag{1}$$

Here, C refers to the channel dimension of the dense features, and $H = {}^{h}/_{p}, W = {}^{w}/_{p}$ with p corresponding to the output stride of the backbone. Based on this image representation, the next step is to decompose the image into semantically meaningful masks to provide a local grouping prior for fitting global class prototypes.

Initial principal mask proposal. To identify the initial principal mask proposal in an image I, we analyze the spatial statistical correlations of its features. Specifically, we consider the empirical feature covariance matrix

$$\Sigma = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\big(f_{:,i,j} - \bar{f}\big)\big(f_{:,i,j} - \bar{f}\big)^{\top}\,, \tag{2}$$

where $f_{:,i,j} \in \mathbb{R}^C$ are the features at position (i, j) and $\bar{f} \in \mathbb{R}^C$ is the mean feature. To identify the feature direction that captures the largest variance in the feature distribution, we seek the first principal component of Σ by solving

$$\Sigma v = \lambda v\,. \tag{3}$$

We obtain the first principal component as the eigenvector v_1 to the largest eigenvalue λ_1, which can be computed efficiently with Singular Value Decomposition (SVD) using the flattened features f.

To identify a candidate region, our next goal is to compute a spatial feature similarity map to the dominant feature direction. We observe that doing so directly with the principal direction does not always lead to sufficient localization, *i. e.*, high similarities arise across multiple visual concepts in an image, elaborated in more detail in Appendix A.1. This can be circumvented by first spatially anchoring the dominant feature vector in the feature map. To that end, we obtain the nearest neighbor feature $\tilde{f} \in \mathbb{R}^C$ of the first principal component v_1 by considering the cosine distance in the normalized feature space $\hat{f}$ as

$$\tilde{f} = \hat{f}_{:,i,j}\,, \quad \text{where} \quad (i, j) = \arg\max_{i,j}\,\big(v_1^{\top}\,\hat{f}\big)\,. \tag{4}$$

Given this, we compute the cosine-similarity map $M \in \mathbb{R}^{H \times W}$ of the dominant feature *w. r. t.* all features as

$$M = (M_{i,j})_{i,j}\,, \quad \text{where} \quad M_{i,j} = \left(\tilde{f}\right)^{\top} \hat{f}_{:,i,j}\,. \tag{5}$$

Next, a threshold $\psi \in (0,1)$ is applied to the similarity map in order to suppress noise and further localize the initial mask. Accordingly, elements of a binary similarity map $P^1 \in \{0,1\}^{H \times W}$ are set to 1 when larger than a fraction ψ of the maximal similarity, and 0 otherwise, *i. e.*,

$$P^1 = \left[M_{i,j} > \psi \cdot \max_{m,n} M_{m,n}\right]_{i,j}\,, \tag{6}$$

where $[\cdot]$ denotes the Iverson bracket. This binary *principal mask* P^1 gives rise to the first principal mask proposal in image I.

Further principal mask proposals. Subsequent mask proposals result from iteratively repeating the described procedure. To that end, it is necessary to suppress features that have already been assigned to a pseudo label. Specifically in iteration z, given the mask proposals P^s, $s = 1, \ldots, z-1$, extracted in previous iterations, we mask out the features that have already been considered as

$$f^z_{:,i,j} = f_{:,i,j} \left[\sum\nolimits_{s=1}^{z-1} P^s_{i,j} = 0\right]\,. \tag{7}$$

Applying Eqs. (2) to (6) on top of the masked features f^z yields principal mask proposal P^z, and so on. We repeat this procedure until the majority of features (*e. g.*, 95%) have been assigned to a mask. In a final step, the remaining features, in case there are any, are assigned to an "ignore" mask

$$P^0_{i,j} = 1 - \sum_{z=1}^{Z-1} P^z_{i,j}\,. \tag{8}$$

This produces a tensor $P \in \{0,1\}^{Z \times H \times W}$ of Z spatial similarity masks decomposing a single image into Z non-overlapping regions.

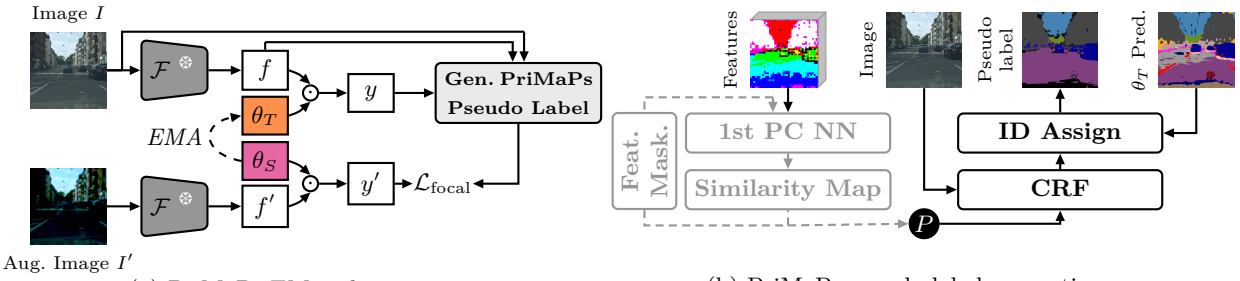
Proposal post-processing. To further improve the alignment of the masks with edges and color-correlated regions in the image, a fully connected Conditional Random Field (CRF) with Gaussian edge potentials (Krähenbühl & Koltun, 2011) is applied to the initial mask proposals P (after bilinear upsampling to the image resolution) for 10 inference iterations.

In order to form a pseudo label for semantic segmentation out of the Z mask proposals, each mask has to be assigned one out of K class labels. This is accomplished using a segmentation prediction of our optimization process, called PriMaPs-EM, detailed below. The entire PriMaPs pseudo label generation process is illustrated in Figure 2b.

3.2 PriMaPs-EM

Shown in Fig. 2, PriMaPs-EM is an iterative optimization technique. It leverages the frozen pre-trained self-supervised backbone model $\mathcal{F}$ and two identically sized vector sets, the class prototypes θ_S and their moving average, the momentum class prototypes θ_T. The class prototypes θ_S and θ_T are the K pseudo class representations in the feature space, projecting the C-dimensional features linearly to K semantic pseudo classes. PriMaPs-EM constructs pseudo labels using PriMaPs, which provide guidance through local consistency for fitting the global class prototypes. In every optimization iteration, we compute the segmentation prediction y from the momentum class prototypes θ_T. Next, we assign the pseudo-class ID that is most frequently predicted within each proposal, yielding the final pseudo-label map $P^* \in \{0,1\}^{K \times h \times w}$, a one-hot encoding of a pseudo-class ID. Finally, we optimize the class prototypes θ_S using the pseudo label.

PriMaPs-EM consists of two stages, since in our case a meaningful initialization of the class prototypes is vital to provide a reasonable optimization signal. This can be traced back to the pseudo-label generation, which utilizes a segmentation prediction to assign globally consistent classes to the masks. Initializing the class prototypes randomly leads to a highly unstable and noisy signal.



(a) PriMaPs-EM architecture

(b) PriMaPs pseudo label generation

Figure 2: **(a) PriMaPs-EM architecture.** Images are embedded by the frozen self-supervised backbone $\mathcal{F}$. First, both class prototypes θ_S and θ_T are initialized via a clustering objective. The segmentation prediction y from the momentum class prototypes θ_T arises via a dot product with the image features f. While PriMaPs are based on f alone, the pseudo labels additionally use the image I and the segmentation prediction y from the momentum class prototypes θ_T. We use the pseudo labels to optimize the class prototypes θ_S, which are gradually transferred to θ_T by means of an EMA. **(b) PriMaPs pseudo label generation.** Masks are proposed by iterative binary partitioning based on the cosine similarity of the features of any unassigned pixel to their first principal component. Next, the masks P are aligned to the image using a CRF (Krähenbühl & Koltun, 2011). Finally, a pseudo-class ID is assigned per mask based on the segmentation prediction from the θ_T. Gray indicates iterative steps.

Initialization. We initialize the class prototypes θ_T with the first K principal components. Next, a cosine distance batch-wise K-means (MacQueen, 1967) loss

$$\mathcal{L}_{K\text{-means}}(\theta_T) = -\sum_{i,j} \max\left(\theta_T^\top f_{:,i,j}\right) \tag{9}$$

is minimized with respect to θ_T for a fixed number of epochs. This minimizes the cumulative cosine distances of the image features $f_{:,i,j}$ to their respective closest class prototype. θ_S is initialized with the same prototypes.

Moving average stochastic EM. In each iteration, we use the backbone features and momentum class prototypes θ_T to yield a segmentation prediction y from which pseudo labels are generated as described in Sec. 3.1. θ_S is optimized by applying a batch-wise focal loss (Lin et al., 2020) with respect to these pseudo labels. The focal loss $\mathcal{L}_{\text{focal}}$ is a weighted version of the cross-entropy loss, increasing the loss contribution of less confident classes, *i. e.*,

$$\mathcal{L}_{\text{focal}}(\theta_S; y') = -\sum_{k,i,j} (1-\chi_k)^2 P^*_{k,i,j} \log(y'_{k,i,j}) \,, \tag{10}$$

where $y'_{:,i,j} = \text{softmax}(\theta_S^\top f_{:,i,j})$ are the predictions and χ_k is the class-wise confidence value approximated by averaging $y'_{:,i,j}$ spatially. The class prototypes θ_S are optimized with an augmented input image I'. We employ photometric augmentations (Gaussian blur, grayscaling, and color jitter), introducing a controlled noise, thereby strengthening the robustness of our class representation. The momentum class prototypes θ_T are the exponential moving average of the class prototypes θ_S. This is utilized in order to stabilize the optimization, accounting for the noisy nature of unsupervised signal used for optimization. We update θ_T every γ_t iterations with a decay γ_ψ as

$$\theta_T^{t+\gamma_t} = \gamma_\psi \theta_T^t + (1-\gamma_\psi)\theta_S^{t+\gamma_t} \,, \tag{11}$$

where t is the iteration index of the previous update. This optimization approach resembles moving average stochastic EM. Hereby, the E-step amounts to finding pseudo labels using PriMaPs and the momentum class prototypes. The M-step optimizes the class prototypes with respect to their focal loss $\mathcal{L}_{\text{focal}}$. Stochasticity arises from performing EM in mini-batches.

Inference. At inference time, we obtain a segmentation prediction from the momentum class prototypes θ_T, refined using a fully connected CRF with Gaussian edge potentials (Krähenbühl & Koltun, 2011) following previous approaches (Van Gansbeke et al., 2021; Hamilton et al., 2022; Seong et al., 2023). This is the identical CRF as already used for refining the masks in the PriMaPs pseudo-label generation. We use the identical CRF parameters as previous work (Van Gansbeke et al., 2021; Hamilton et al., 2022; Seong et al., 2023).

4 Experiments

To assess the efficacy of our approach, we compare it to the current state-of-the-art in unsupervised semantic segmentation. For a fair comparison, we closely follow the overall setup used by numerous previous works (Ji et al., 2019; Cho et al., 2021; Hamilton et al., 2022; Seong et al., 2023).

4.1 Experimental Setup

Datasets. Following the practice of previous work, we conduct experiments on Cityscapes (Cordts et al., 2016), COCO-Stuff (Caesar et al., 2018), and Potsdam-3 (ISPRS). Cityscapes and COCO-Stuff are evaluated using 27 classes, while Potsdam is evaluated on the 3-class variant. Adopting the established evaluation protocol (Ji et al., 2019; Cho et al., 2021; Hamilton et al., 2022; Seong et al., 2023), we resize images to 320 pixels along the smaller axis and crop the center 320×320 pixels. This is adjusted to 322 pixels for DINOv2. Different from previous work, we apply this simple scheme throughout this work, thus dispensing with elaborate multi-crop approaches of previous methods (Hamilton et al., 2022; Yin et al., 2022; Seong et al., 2023).

Self-supervised backbone. Experiments are conducted across a collection of pre-trained self-supervised feature embeddings: DINO (Caron et al., 2021) based on ViT-Small and ViT-Base using 8×8 patches; and DINOv2 (Oquab et al., 2024) based on ViT-Small and ViT-Base using 14×14 patches. In the spirit of SSL principles, we keep the backbone parameters frozen throughout the experiments. We use the output from the last network layer as our SSL feature embeddings. Since PriMaPs-EM is agnostic to the used embedding space, we can also apply it on top of current state-of-the-art unsupervised segmentation pipelines. Here, we consider STEGO (Hamilton et al., 2022) and HP (Seong et al., 2023), which also use DINO features but learn a target domain-specific subspace.

Baseline. Following (Hamilton et al., 2022; Seong et al., 2023), we train a single linear layer as a baseline with the same structure as θ_S and θ_T by minimizing the cosine distance batch-wise K-Means loss from Eq. (9). Hereby, parameters, such as the number of epochs and the learning rate, are identical to those used when employing PriMaPs-EM.

PriMaPs-EM. As discussed in Sec. 3.2, the momentum class prototypes θ_T are initialized using the first K principal components; we use 2975 images for PCA, as this is the largest number of training images shared by all datasets. Next, θ_T is pre-trained by minimizing Eq. (9) using Adam (Kingma & Ba, 2015). We use a learning rate of 0.005 for 2 epochs on all datasets and backbones. The weights are then copied to θ_S. For fitting the class prototypes using EM, θ_S is optimized by minimizing the focal loss from Eq. (10) with Adam (Kingma & Ba, 2015) using a learning rate of 0.005. The momentum class prototypes θ_T are updated using an EMA according to Eq. (11) every $\gamma_s = 10$ steps with decay $\gamma_\psi = 0.98$. We set the PriMaPs mask proposal threshold to $\psi = 0.4$. We use a batch size of 32 for 50 epochs on Cityscapes and Potsdam-3, and use 5 epochs on COCO-Stuff due to its larger size. Importantly, the same hyperparameters are used across *all* datasets and backbones. Moreover, note that fitting class prototypes with PriMaPs-EM is quite practical, *e. g.*, about 2 hours on Cityscapes. Experiments are conducted on a single NVIDIA A6000 GPU.

Supervised upper bounds. To assess the potential of the SSL features used, we report supervised upper bounds. Specifically, we train a linear layer using cross-entropy and Adam with a learning rate of 0.005. Since PriMaPs-EM uses frozen SSL features, its supervised bound is the same as that of the underlying features. This is not the case, however, for prior work (Hamilton et al., 2022; Seong et al., 2023), which project the feature representation affecting the upper bound.

Table 1: **Cityscapes – PriMaPs-EM *(Ours)* comparison to existing unsupervised semantic segmentation methods**, using Accuracy and mean IoU (in %) for unsupervised and supervised probing. Double citations refer to a method's origin and the work conducting the experiment.

Method	Backbone	*Unsupervised*		*Supervised*	
		Acc	**mIoU**	**Acc**	**mIoU**
IIC (Ji et al., 2019; Cho et al., 2021)	ResNet18 +FPN	47.9	6.4	–	–
MDC (Caron et al., 2018; Cho et al., 2021)		40.7	7.1	–	–
PiCIE (Cho et al., 2021)		65.5	12.3	–	–
VICE (Karlsson et al., 2022)		31.9	12.8	86.3	31.6
Baseline (Caron et al., 2021)	DINO ViT-S/8	61.4	15.8	91.0	35.4
+ TransFGU (Yin et al., 2022)		77.9	16.8	–	–
+ HP (Seong et al., 2023)		80.1	18.4	91.2	30.6
+ PriMaPs-EM		**81.2**	**19.4**	91.0	35.4
+ HP (Seong et al., 2023) + PriMaPs-EM		76.6	19.2	91.2	30.6
Baseline (Caron et al., 2021)	DINO ViT-B/8	49.2	15.5	91.6	35.9
+ STEGO (Hamilton et al., 2022; Koenig et al., 2023)		73.2	21.0	89.6	28.0
+ HP (Seong et al., 2023)		**79.5**	18.4	90.9	33.0
+ PriMaPs-EM		59.6	17.6	91.6	35.9
+ STEGO (Hamilton et al., 2022) + PriMaPs-EM		78.6	**21.6**	89.6	28.0
Baseline (Oquab et al., 2024)	DINOv2 ViT-S/14	49.5	15.3	90.8	41.9
+ PriMaPs-EM		**71.5**	**19.0**	90.8	41.9
Baseline (Oquab et al., 2024)	DINOv2 ViT-B/14	36.1	14.9	91.0	44.8
+ PriMaPs-EM		**82.9**	**21.3**	91.0	44.8

Evaluation. For inference, we use the prediction from the momentum class prototypes θ_T. CRF refinement uses 10 inference iterations and standard parameters $a = 4, b = 3, \theta_\alpha = 67, \theta_\beta = 3, \theta_\gamma = 1$ from prior work (Van Gansbeke et al., 2021; Hamilton et al., 2022; Seong et al., 2023). We evaluate common metrics in unsupervised semantic segmentation, specifically the mean Intersection over Union (mIoU) and Accuracy (Acc) over all classes after aligning the predicted class IDs with ground-truth labels by means of Hungarian matching (Kuhn, 1955).

SotA + PriMaPs-EM. To explore our method's potential, we additionally employ PriMaPs-EM on top of STEGO (Hamilton et al., 2022) and HP (Seong et al., 2023). For each backbone-dataset combination, we apply it on top of the best previous method in terms of mIoU. To that end, the training signal for learning the feature projection of (Hamilton et al., 2022; Seong et al., 2023) remains unchanged. We apply PriMaPs-EM fully orthogonally, using the DINO backbone features for pseudo-label generation and fit a direct connection between the feature space of the state-of-the-art method and the prediction space.

4.2 Results

We compare PriMaPs-EM against prior work for unsupervised semantic segmentation (Ji et al., 2019; Cho et al., 2021; Hamilton et al., 2022; Yin et al., 2022; Li et al., 2023; Seong et al., 2023). As in previous work, we use DINO (Caron et al., 2021) as the main baseline. Additionally, we also test PriMaPs-EM on top of DINOv2 (Oquab et al., 2024), STEGO (Hamilton et al., 2022), and HP (Seong et al., 2023). Overall, we observe that the DINO baseline already achieves strong results (*cf.* Tabs. 1 to 3). DINOv2 features significantly raise the supervised upper bounds in terms of Acc and mIoU, the improvement in the unsupervised case remains more modest. Nevertheless, PriMaPs-EM further boosts the unsupervised segmentation performance.

In Tab. 1, we compare to previous work on the Cityscapes dataset. PriMaPs-EM leads to a consistent improvement over all baselines in terms of unsupervised segmentation accuracy. For example, PriMaPs-EM boosts DINO ViT-S/8 by +3.6% and +19.8% in terms of mIoU and Acc, respectively, which leads to

Table 2: **COCO-Stuff – PriMaPs-EM *(Ours)* comparison to existing unsupervised semantic segmentation methods**, using Accuracy and mean IoU (in %) for unsupervised and supervised probing. Double citations refer to a method's origin and the work conducting the experiment.

Method	Backbone	*Unsupervised*		*Supervised*	
		Acc	**mIoU**	**Acc**	**mIoU**
IIC (Ji et al., 2019; Cho et al., 2021)	ResNet18 +FPN	21.8	6.7	44.5	8.4
MDC (Caron et al., 2018; Cho et al., 2021)		32.2	9.8	48.6	13.3
PiCIE (Cho et al., 2021)		48.1	13.8	54.2	13.9
PiCIE+H (Cho et al., 2021)		50.0	14.4	54.8	14.8
VICE (Karlsson et al., 2022)		28.9	11.4	62.8	25.5
Baseline (Caron et al., 2021)	DINO ViT-S/8	34.2	9.5	72.0	41.3
+ TransFGU (Yin et al., 2022)		52.7	17.5	–	–
+ STEGO (Hamilton et al., 2022)		48.3	24.5	74.4	38.3
+ ACSeg (Li et al., 2023)		–	16.4	–	–
+ HP (Seong et al., 2023)		57.2	24.6	75.6	42.7
+ PriMaPs-EM		46.5	16.4	72.0	41.3
+ HP (Seong et al., 2023) + PriMaPs-EM		**57.8**	**25.1**	75.6	42.7
Baseline (Caron et al., 2021)	DINO ViT-B/8	38.8	15.7	74.0	44.6
+ STEGO (Hamilton et al., 2022)		56.9	28.2	76.1	41.0
+ PriMaPs-EM		48.5	21.9	74.0	44.6
+ STEGO (Hamilton et al., 2022) + PriMaPs-EM		**57.9**	**29.7**	76.1	41.0
Baseline (Oquab et al., 2024)	DINOv2 ViT-S/14	44.5	22.9	77.9	52.8
+ PriMaPs-EM		**46.5**	**23.8**	77.9	52.8
Baseline (Oquab et al., 2024)	DINOv2 ViT-B/14	35.0	17.9	77.3	53.7
+ PriMaPs-EM		**52.8**	**23.6**	77.3	53.7

state-of-the-art performance. Notably, we find PriMaPs-EM to be complementary to other state-of-the-art unsupervised segmentation methods like STEGO (Hamilton et al., 2022) and HP (Seong et al., 2023) on the corresponding backbone model. This suggests that these methods use their SSL representation only to a limited extent and do not fully leverage the inherent properties of the underlying SSL embeddings. Similar observations can be drawn for the experiments on COCO-Stuff in Tab. 2. PriMaPs-EM leads to a consistent improvement across all four SSL baselines, as well as an improvement over STEGO and HP. For instance, combining STEGO with PriMaPs-EM leads to +14.0% and +19.1% improvement over the baseline in terms of mIoU and Acc for DINO ViT-B/8. Experiments on the Potsdam-3 dataset follow the same pattern (*cf.* Tab. 3). PriMaPs-EM leads to a consistent gain over the baseline, *e. g.* +17.6% and +14.4% in terms of mIoU and Acc, respectively, for DINO ViT-B/8. Moreover, it also boosts the accuracy of STEGO and HP. In some cases, the gain of PriMaPs-EM is limited. For example, in Tab. 1 for DINO ViT-B/8 + PriMaPs-EM, the class prototype for "sidewalk" is poor while the classes "road" and "vegetation" superimpose smaller objects. For DINO ViT-S/8 + PriMaPs-EM in Tab. 3, the class prototype "road" is poor. This limits the overall performance of our method while still outperforming the respective baseline in both cases.

Overall, PriMaPs-EM provides modest but consistent benefits over a wide range of baselines and datasets and reaches competitive segmentation performance *w. r. t.* the state-of-the-art using identical hyperparameters across *all* backbones and datasets. Recalling the simplicity of the techniques behind PriMaPs, we believe that this is a significant result. The complementary effect of PriMaPs-EM on other state-of-the-art methods (STEGO, HP) further suggests that they rely on DINO features for mere "bootstrapping" and learn feature representations with orthogonal properties to those of DINO. We conclude that PriMaPs-EM constitutes a straightforward, entirely orthogonal tool for boosting unsupervised semantic segmentation.

4.3 Ablation Study

To untangle the factors behind PriMaPs-EM, we examine the individual components in a variety of ablation experiments to access the contribution.

Table 3: **Potsdam-3 – PriMaPs-EM *(Ours)* comparison to existing unsupervised semantic segmentation methods**, using Accuracy and mean IoU (in %) for unsupervised and supervised probing. Double citations refer to a method's origin and the work conducting the experiment.

Method	Backbone	*Unsupervised*		*Supervised*	
		Acc	**mIoU**	**Acc**	**mIoU**
RandomCNN (Cho et al., 2021)	VGG 11	38.2	–	–	–
K-Means (Pedregosa et al., 2011; Cho et al., 2021)		45.7	–	–	–
SIFT (Lowe, 2004; Cho et al., 2021)		38.2	–	–	–
ContextPrediction (Doersch et al., 2015; Cho et al., 2021)		49.6	–	–	–
CC (Isola et al., 2015; Cho et al., 2021)		63.9	–	–	–
DeepCluster (Caron et al., 2018; Cho et al., 2021)		41.7	–	–	–
IIC (Ji et al., 2019; Cho et al., 2021)		65.1	–	–	–
Baseline (Caron et al., 2021)	DINO ViT-S/8	56.6	33.6	82.0	69.0
+ STEGO (Hamilton et al., 2022; Koenig et al., 2023)		77.0	62.6	85.9	74.8
+ PriMaPs-EM		62.5	38.9	82.0	69.0
+ STEGO (Hamilton et al., 2022) + PriMaPs-EM		**78.4**	**64.2**	85.9	74.8
Baseline (Caron et al., 2021)	DINO ViT-B/8	66.1	49.4	84.3	72.8
+ HP (Seong et al., 2023)		82.4	69.1	88.0	78.4
+ PriMaPs-EM		80.5	67.0	84.3	72.8
+ HP (Seong et al., 2023)+ PriMaPs-EM		**83.3**	**71.0**	88.0	78.4
Baseline (Oquab et al., 2024)	DINOv2 ViT-S/14	75.9	61.0	86.6	76.2
+ PriMaPs-EM		**78.5**	**64.3**	86.6	76.2
Baseline (Oquab et al., 2024)	DINOv2 ViT-B/14	82.4	69.9	87.9	78.3
+ PriMaPs-EM		**83.2**	**71.1**	87.9	78.3

Table 4: **Ablation study** analyzing design choices and components in the PriMaPs pseudo-label generation *(a)* and PriMaPs-EM *(b)* for COCO-Stuff using DINO ViT-B/8.

(a) PriMaPs pseudo label ablation

Method	**Acc**	**mIoU**
Baseline (Caron et al., 2021)	38.8	15.7
Similarity Masks	46.3	19.8
+ NN	44.9	20.0
+ P-CRF (≡ PriMaPs-EM)	48.4	21.9
PriMaPs-EM (non-iter.)	47.9	21.7

(b) PriMaPs-EM ablation

Method	**Acc**	**mIoU**
Baseline (Caron et al., 2021)	38.8	15.7
+ PriMaPs pseudo label	38.8	18.0
+ EMA	45.0	20.2
+ Augment	46.0	20.4
+ CRF (≡ PriMaPs-EM)	48.4	21.9

PriMaPs pseudo-label ablations. In Tab. 4a, we analyze the contribution of the individual sub-steps for PriMaPs pseudo-label generation by increasing the complexity of label generation. We provide the DINO baseline, which corresponds to K-means feature clustering, for reference. In the most simplified case, we directly use the similarity mask, similar to Eq. (4). Next, we use the nearest neighbor (+NN in Tab. 4a) of the principal component to get the masks as in Eq. (5), followed by the full approach with CRF refinement (+P-CRF). Except for the changes in the pseudo-label generation, the optimization remains as described in Sec. 4.1. We observe that the similarity masks already provide a good staring point, yet we identify a gain from every single component step. This suggests that using the nearest neighbor improves the localization of the similarity mask. Similarly, CRF refinement improves the alignment between the masks and the image content. We also experiment with using the respective next principal direction (non-iter.) instead of iteratively extracting the first component from masked features. This leads to slightly inferior results.

PriMaPs-EM architecture ablations. In a similar vein, we analyze the contribution of the different architectural components of PriMaPs-EM. Optimizing over a single set of class prototypes using the proposed

Table 5: **Oracle quality assessment of PriMaPs pseudo labels** for Cityscapes, COCO-Stuff, and Potsdam-3 by assigning oracle class IDs to the masks. "Pseudo" refers to evaluating only the pixels contained in the pseudo label, "All" to evaluating including the "ignore" assignments of the pseudo label.

Method	*Cityscapes*		*COCO-Stuff*		*Potsdam-3*	
	Acc	**mIoU**	**Acc**	**mIoU**	**Acc**	**mIoU**
Pseudo	92.4	54.0	93.4	82.4	95.2	90.9
All	73.2	32.4	74.1	55.9	67.4	48.9
DINO ViT-B/8 Baseline (Caron et al., 2021)	49.2	15.5	38.8	15.7	66.1	49.4

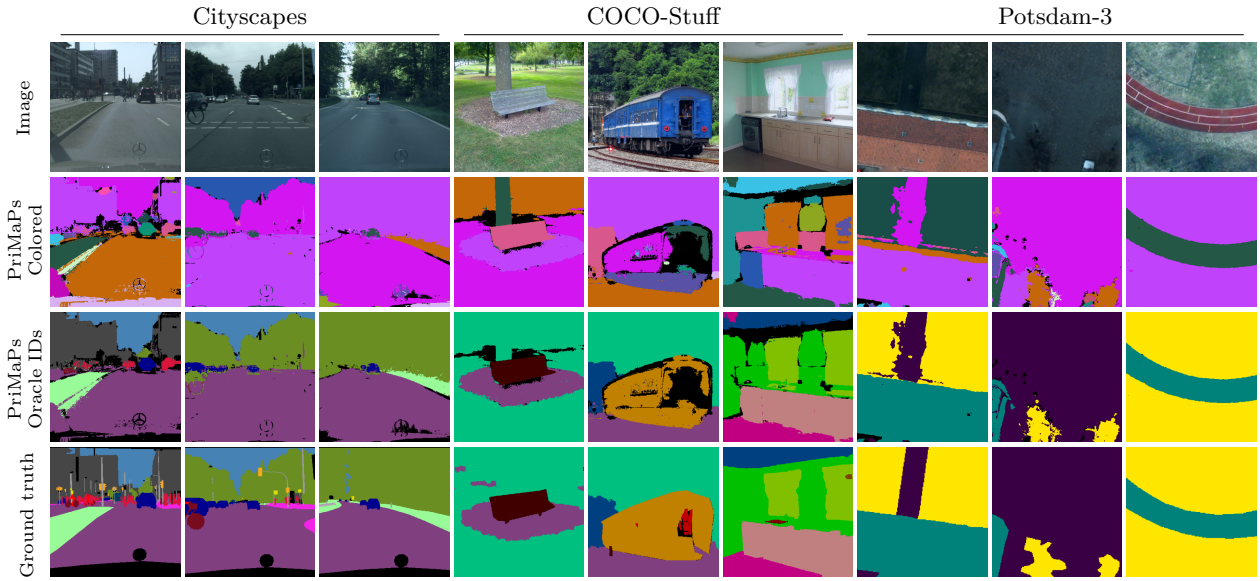


Figure 3: **Qualitative PriMaPs examples** using DINO ViT-B/8 for Cityscapes, COCO-Stuff, and Potsdam-3. *PriMaPs Colored* – each mask proposal is visualized in a different color. *PriMaPs Oracle class IDs* – each mask is colored in the corresponding ground-truth class color.

PriMaPs pseudo labels already provides moderate improvement (+PriMaPs pseudo label in Tab. 4b), despite the disadvantage of an unstable and noisy optimization signal. Adding the EMA (+EMA) leads to a more stable optimization and further improved segmentation. Augmenting the input (+Augment) results in a further gradual improvement. Similarly, refining the prediction with a CRF improves the results further (+CRF).

Assessing PriMaPs pseudo labels. To estimate the quality of the pseudo labels, respectively the principal masks, we decouple those from the class ID assignment by providing the oracle ground-truth class for each mask in Tab. 5. To that end, we evaluate all pixels included in our pseudo labels ("Pseudo"), corresponding to the upper bound of our optimization signal. Furthermore, we evaluate "All" by assigning the "ignore" pixels to a wrong class. The results indicate a high quality of the pseudo-label maps. Fig. 3 shows qualitative examples of the PriMaPs mask proposals and pseudo labels. We visualize individual masks, each in a different color (*PriMaPs Colored*). We also display oracle pseudo labels assigning each mask a color based on the ground-truth label (*PriMaPs Oracle class IDs*). We observe that the mask proposals align well with the ground-truth labels across all three datasets, generalizing across three distinct domains. PriMaPs effectively partitions images into semantically meaningful masks.

Qualitative results. We show qualitative results for Cityscapes, COCO-Stuff, and Potsdam-3 in Fig. 4. We observe that PriMaPs-EM leads to less noisy results compared to the baseline, showcasing an improved

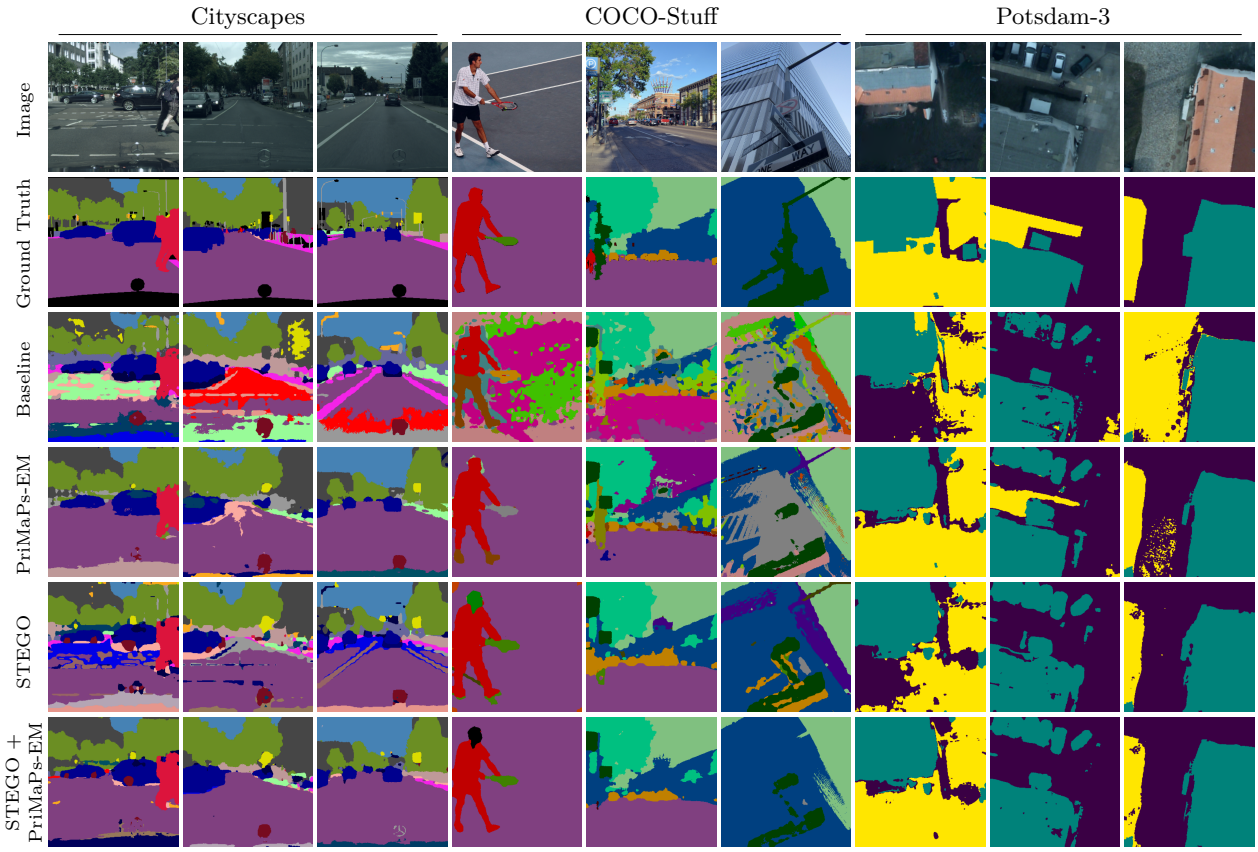


Figure 4: **Qualitative results** for the DINO ViT-B/8 baseline, PriMaPs-EM *(Ours)*, STEGO (Hamilton et al., 2022), and STEGO+PriMaPs-EM *(Ours)* for Cityscapes, COCO-Stuff, and Potsdam-3. Our method produces locally more consistent segmentation results reducing overall misclassification compared to the corresponding baseline.

local consistency of the segmentation and reduced mis-classification. The comparison with STEGO as a baseline exhibits a similar trend. For further examples and comparisons with HP, please refer to Appendix B.2.

Limitations. One of the main challenges is to distinguish between classes that happen to share the same SSL feature representation. This is hardly avoidable if the feature representation is fixed, as was the case here and in previous work (Hamilton et al., 2022; Seong et al., 2023). Another limitation across existing unsupervised semantic segmentation approaches is the limited spatial image resolution. This limitation comes from the SSL training objectives (Caron et al., 2021; Oquab et al., 2024), which are image-level, rather than pixel-level. As a result, we can observe difficulties in segmenting very small, finely resolved structures.

5 Conclusion

We present PriMaPs, a novel dense pseudo-label generation approach for unsupervised semantic segmentation. We derive lightweight mask proposals directly from off-the-shelf self-supervised learned features, leveraging the intrinsic properties of their embedding space. Our mask proposals can be used as pseudo labels to effectively fit global class prototypes using moving average stochastic EM with PriMaPs-EM. Despite the simplicity, PriMaPs-EM leads to a consistent boost in unsupervised segmentation accuracy when applied to a variety of SSL features or orthogonally to current state-of-the-art unsupervised semantic segmentation pipelines, as shown by our results across multiple datasets.

References

Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep ViT features as dense visual descriptors. In *ECCVW*, volume 13804, pp. 39–55, 2022.

Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In *CVPR*, pp. 4253–4262, 2020.

Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE T. Pattern Anal. Mach. Intell.*, 33(5):898–916, 2011.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, and Simran Arora et al. On the opportunities and risks of foundation models. *arXiv:2108.07258 [cs.LG]*, 2021.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-Stuff: Thing and stuff classes in context. In *CVPR*, pp. 1209–1218, 2018.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, volume 11218, pp. 132–149, 2018.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*2020*, volume 33, pp. 9912–9924, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pp. 9650–9660, 2021.

Jianfei Chen, Jun Zhu, Yee Whye Teh, and Tong Zhang. Stochastic expectation maximization with variance reduction. In *NeurIPS*2018*, volume 31, pp. 7967–7977, 2018a.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE T. Pattern Anal. Mach. Intell.*, 40(4):834–848, 2018b.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pp. 1597–1607, 2020.

Jang Hyun Cho, Utkarsh Mall, Kavita Bala, and Bharath Hariharan. PiCIE: Unsupervised semantic segmentation using invariance and equivariance in clustering. In *CVPR*, pp. 16794–16804, 2021.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Scharwächter, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *CVPR*, pp. 3213–3223, 2016.

Jifeng Dai, Kaiming He, and Jian Sun. BoxSup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *ICCV*, pp. 1635–1643, 2015.

Chris Ding and Xiaofeng He. K-means clustering via principal component analysis. In *ICML*, 2004.

Carl Doersch, Abhinav Kumar Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, pp. 1422–1430, 2015.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020.

Petros Drineas, Alan M. Frieze, Ravi Kannan, Santosh S. Vempala, and V. Vinay. Clustering large graphs via the singular value decomposition. *Machine Learning*, 56(1-3):9–33, 2004.

Linus Ericsson, Henry Gouk, and Timothy M. Hospedales. How well do self-supervised models transfer? In *CVPR*, pp. 5414–5423, 2021.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent – A new approach to self-supervised learning. In *NeurIPS*2020*, volume 33, pp. 21271–21284, 2020.

Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T. Freeman. Unsupervised semantic segmentation by distilling feature correspondences. In *ICLR*, 2022.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pp. 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pp. 16000–16009, 2022.

Jyh-Jing Hwang, Stella X. Yu, Jianbo Shi, Maxwell D. Collins, Tien-Ju Yang, Xiao Zhang, and Liang-Chieh Chen. SegSort: Segmentation by discriminative sorting of segments. In *ICCV*, pp. 7333–7343. IEEE, 2019.

Phillip Isola, Daniel Zoran, Dilip Krishnan, and Edward H Adelson. Learning visual groups from co-occurrences in space and time. In *arXiv:1511.06811 [cs.CV]*, 2015.

ISPRS. ISPRS 2d semantic labeling contest. `https://www.isprs.org/education/benchmarks/UrbanSemLab` Accessed: 2024-04-19.

Xu Ji, Joao F. Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *ICCV*, pp. 9865–9874, 2019.

Robin Karlsson, Tomoki Hayashi, Keisuke Fujii, Alexander Carballo, Kento Ohtani, and Kazuya Takeda. VICE: Improving dense representation learning by superpixelization and contrasting cluster assignment. In *BMVC*, 2022.

Tsung-Wei Ke, Jyh-Jing Hwang, Yunhui Guo, Xudong Wang, and Stella X. Yu. Unsupervised hierarchical semantic segmentation with multiview cosegmentation and clustering transformers. In *CVPR*, pp. 2561–2571, 2022.

Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *ICCV*, 2023.

Alexander Koenig, Maximilian Schambach, and Johannes Otterbach. Uncovering the inner workings of STEGO for safe unsupervised semantic segmentation. In *CVPRW*, 2023.

Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected CRFs with Gaussian edge potentials. In *NIPS*2011*, volume 24, 2011.

Harold W. Kuhn. The Hungarian method for the assignment problem. In *NVL*, volume 52, 1955.

Kehan Li, Zhennan Wang, Zesen Cheng, Runyi Yu, Yian Zhao, Guoli Song, Li Yuan, and Jie Chen. Dynamic clustering network for unsupervised semantic segmentation. In *CVPR*, pp. 7162–7172, 2023.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE T. Pattern Anal. Mach. Intell.*, 42(2):318–327, 2020.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pp. 3431–3440, 2015.

David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2): 91–110, 2004.

James MacQueen. Some methods for classification and analysis of multivariate observations. In *Berkeley Symp. on Math. Statist. and Prob.*, volume 1, pp. 281–297. Oakland, CA, USA, 1967.

Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE T. Pattern Anal. Mach. Intell.*, 44(7):3523–3542, 2022.

Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. In *NeurIPS*2021*, pp. 23296–23308, 2021.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, volume 9910, pp. 69–84, 2016.

Youngmin Oh, Beomjun Kim, and Bumsub Ham. Background-aware pooling and noise-aware loss for weakly-supervised semantic segmentation. In *CVPR*, pp. 6909–6918, 2021.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. In *TMLR*, 2024.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *NeurIPS*2019*, pp. 8024–8035, 2019.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, pp. 2536–2544, 2016.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. In *JMLR*, volume 12, pp. 2825–2830, 2011.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, volume 9351, pp. 234–241, 2015.

Lixiang Ru, Heliang Zheng, Yibing Zhan, and Bo Du. Token contrast for weakly-supervised semantic segmentation. In *CVPR*, pp. 3093–3102, 2023.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vision*, 115(13):211–252, 2015.

Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, and Francesco Locatello. Bridging the gap to real-world object-centric learning. In *ICLR*, 2023.

Hyun Seok Seong, WonJun Moon, SuBeen Lee, and Jae-Pil Heo. Leveraging hidden positives for unsupervised semantic segmentation. In *CVPR*, pp. 19540–19549, 2023.

Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE T. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000.

Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing ViT features for semantic appearance transfer. In *CVPR*, pp. 10738–10747, 2022.

Jasper R. R. Uijlings, Koen E. A. van de Sande, Theo Gevers, and Arnold W. M. Smeulders. Selective search for object recognition. *Int. J. Comput. Vision*, 104(2):154–171, 2013.

Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Unsupervised semantic segmentation by contrasting object mask proposals. In *ICCV*, pp. 10052–10062, 2021.

Wouter Van Gansbeke, Simon Vandenhende, and Luc Van Gool. Discovering object masks with transformers for unsupervised semantic segmentation. In *arXiv:2206.06363 [cs.CV]*, 2022.

Xudong Wang, Rohit Girdhar, Stella X Yu, and Ishan Misra. Cut and learn for unsupervised object detection and instance segmentation. In *CVPR*, pp. 3124–3134, 2023a.

Yangtao Wang, Xi Shen, Yuan Yuan, Yuming Du, Maomao Li, Shell Xu Hu, James L. Crowley, and Dominique Vaufreydaz. TokenCut: Segmenting objects in images and videos with self-supervised transformer and normalized cut. In *IEEE T. Pattern Anal. Mach. Intell.*, volume 45, pp. 15790–15801, 2023b.

Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaïd, and Dan Xu. Multi-class token transformer for weakly supervised semantic segmentation. In *CVPR*, pp. 4300–4309, 2022.

Zhaoyuan Yin, Pichao Wang, Fan Wang, Xianzhe Xu, Hanling Zhang, Hao Li, and Rong Jin. TransFGU: A top-down approach to fine-grained unsupervised semantic segmentation. In *ECCV*, volume 13689, pp. 73–89, 2022.

Andrii Zadaianchuk, Matthäus Kleindessner, Yi Zhu, Francesco Locatello, and Thomas Brox. Unsupervised semantic segmentation with self-supervised object-centric representations. In *ICLR*, 2023.

A Further Analysis

In this appendix, we provide more detailed insights into PriMaPs-EM beyond the scope of the main paper.

A.1 Nearest Neighbor Anchoring

As described in Sec. 3 of the main paper, PriMaPs are spatially anchoring the first principal component in each iteration of the mask proposal generation. Corresponding to the quantitative findings (*cf.* Tab. 4a), here we additionally analyze this qualitatively. Fig. 5 shows similarity maps of all features to the iteratively computed first principal component (Similarity 1st PC) as well as to the respective nearest neighbor feature (Similarity 1st PC NN) for example images from Cityscapes (Cordts et al., 2016), COCO-Stuff (Caesar et al., 2018), and Potsdam-3 (ISPRS). We observe that the originally computed principal direction can have high similarities with multiple semantic concepts in an image. Hence, finding a suitable threshold that isolates a single main concept is difficult.

However, using the nearest image feature to the principal component as an anchoring element helps to circumvent high similarity values to multiple visual concepts. For instance, in the first example for Cityscapes, the similarity map for the first principal component has high similarities to both "building" and "vegetation". In contrast, the similarity map for the next nearest neighbor of the first principal component results in high similarity to the class "vegetation" only. Consequently, a suitable mask proposal can be obtained through thresholding. We observe this particularly for the Cityscapes dataset and some examples of COCO-Stuff. Further, in cases where the similarity map is already localized to one visual concept, spatial anchoring merely leads to a change in the similarity values without changing the shape of the thresholded proposal.

A.2 Ablation of the Threshold

In the spirit of unsupervised learning, our method effectively has only a single additional hyperparameter – the threshold ψ. Furthermore, this parameter can be set simply by examining the mask proposal as detailed next. In addition, it should be noted that we keep this parameter unchanged for all backbone models and datasets, which further emphasizes the generalizability of our method. As described in Sec. 3 of the main paper, the threshold ψ is used to remove noise in the similarity masks and to localize the optimization signal. In Fig. 6, qualitative examples of PriMaPs mask proposals are visualized for DINO ViT-B/8 on Cityscapes, COCO-Stuff, and Potsdam-3. We show mask proposals for $\psi = 0.3$, $\psi = 0.4$, and $\psi = 0.5$. Visually, it can be observed that meaningful masks are produced for all thresholds. Especially those for threshold 0.3 and 0.4 align very well with the semantic content in the images. While $\psi = 0.3$ seems to provide better mask proposals for COCO-Stuff and Potsdam-3, a problem arises with Cityscapes. Here, the mask proposals contain several semantic concepts and spatially small objects in a large mask (*cf.* the second Cityscapes example, where the mask of the bush in the left half of the image covers both the traffic sign in the foreground, parts of the house in the background, as well as the sky). Since this would lead to a poor optimization signal and the masks of the other two datasets using $\psi = 0.4$ seem visually appealing, the threshold is set to 0.4 in all other experiments.

To further shed light on these qualitative observations, we apply PriMaPs-EM for the scenario described above and vary the threshold from $\psi = 0.2$ to $\psi = 0.6$ with a step size of 0.1. We perform this with the DINO ViT-B/8 backbone for Cityscapes, COCO-Stuff, and Potsdam-3 and show the mIoU in Fig. 7. The quantitative results reflect our qualitative conclusion of the threshold well, and show the trade-off between better segmentation accuracy on COCO-Stuff and Potsdam-3 for a lower threshold and vice versa for Cityscapes. In numbers, a threshold of $\psi = 0.5$ instead of $\psi = 0.4$ for Cityscapes would lead to an additional gain of 1.4 % mIoU, but also to slight losses on COCO-Stuff of 0.8 %. Additionally, this results in more mask proposal iterations. We conclude that the determined threshold $\psi = 0.4$ appears to be reasonable. Even if better results could be achieved for some backbone dataset combinations with individually set thresholds, we consider a fixed threshold that generalizes well across all scenarios to be sensible. We would like to emphasize that this experiment was conducted solely for this single backbone and serves only to validate the qualitative judgement from above. Importantly, we did not determine the hyperparameters based on the evaluation sets.

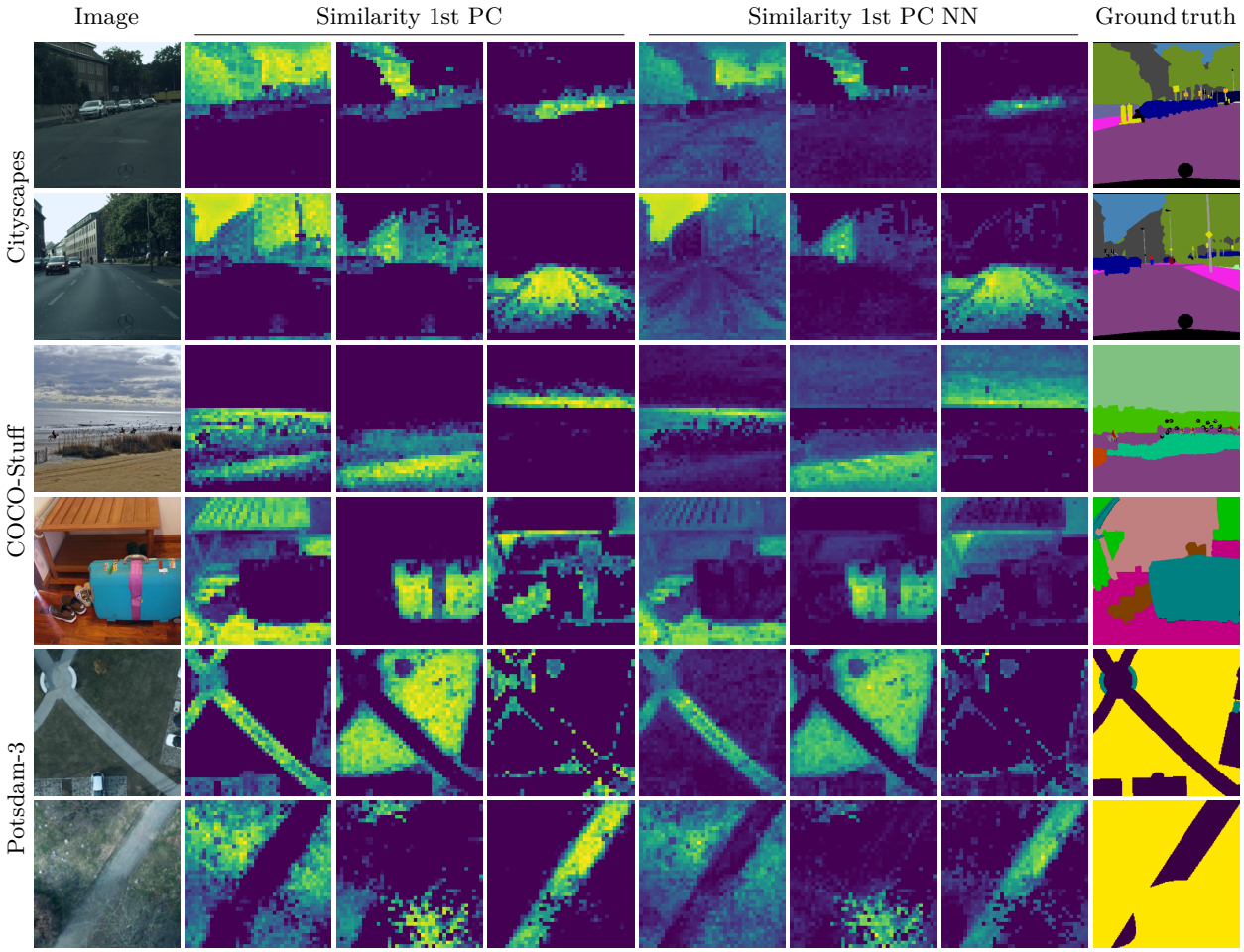


Figure 5: **Nearest neighbor spatial anchoring of the principal direction in PriMaPs**. Image, ground-truth label, and the first three similarity maps with respect to the principal direction *(left)* and their nearest neighbor *(right)* for all three datasets using DINO ViT-B/8. Anchoring localizes the signal for principal directions with high similarities to multiple visual concepts.

B Further Experiments

This appendix provides insights beyond the experiments and ablations shown in Sec. 4.

B.1 Class-Level Quantitative Analysis

To gain a deeper understanding of PriMaPs-EM, we assess the segmentation accuracy in terms of IoU for individual classes. Additionally, we present the confusion matrices among the semantic classes for the DINO ViT-B/8 Baseline, PriMaPs-EM, STEGO, and STEGO+PriMaPs-EM for the COCO-Stuff dataset in Fig. 8. Generally, we can observe that for both the DINO and STEGO baseline, for certain classes (*e. g.*, "Appliance", "Indoor", "Kitchen") the discovered unsupervised class concept does not correlate with human-defined semantic classes. This suggests that the respective backbone feature representation may already be hard to separate. Furthermore, some of the 27 intermediate COCO-Stuff classes merge visually distinct concepts. For instance, the class "indoor" combines "hairbrush", "toothbrush", "hair dryer", "teddy bear", "scissors", "vase", "clock", and "book". We see this assumption partially confirmed by analyzing the class IoUs of linear probing of the DINO features. For the problematic classes, the linear probing IoUs are in the range of approx. 16 %–30 %, whereas for the other classes the IoU is 50 % and higher. We conclude that if it

Figure 6: **Qualitative threshold ψ ablation** with DINO ViT-B/8 using different ψ values for PriMaPs mask proposal generation on all three datasets. The hyperparameter exhibits favorable stability properties.

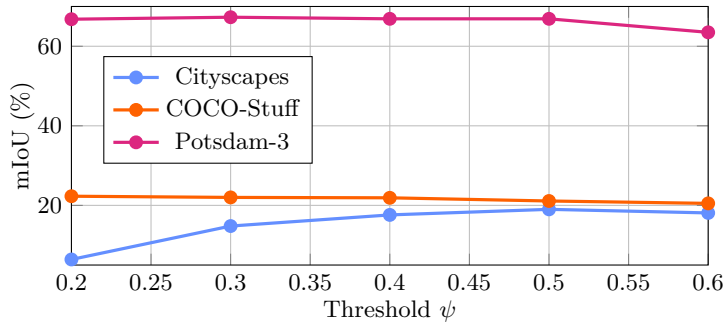


Figure 7: **Quantitative threshold ψ ablation** with DINO ViT-B/8 using different ψ values for PriMaPs mask proposal generation on all three datasets. The single hyperparameter of our method exhibits favorable stability properties.

is already difficult to linearly distinguish the classes based on the backbone features, our method can hardly improve upon this.

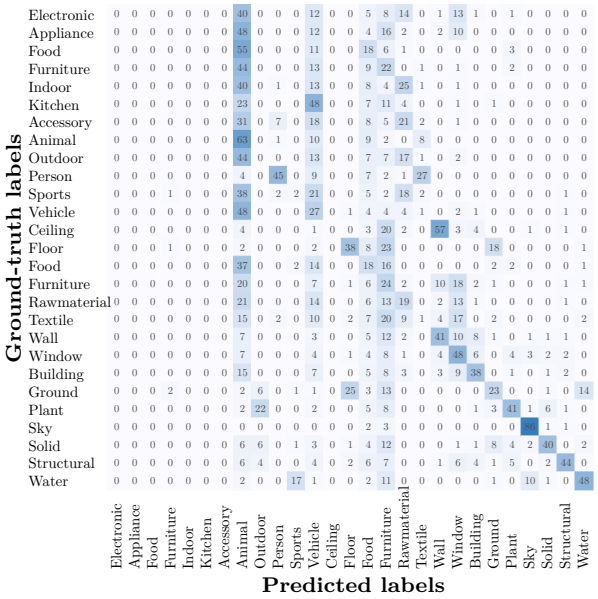
However, for classes meeting this requirement, our method clearly boosts class IoUs. In rare cases (*e. g.*, STEGO comparison to STEGO+PriMaPs-EM for classes "Outdoor" and "Sports"), there is a decrease in one class IoU with PriMaPs-EM while another class IoU increases. This can occur if the change in the prototype representation results in a change of the Hungarian matching for evaluation, though this is rarely observed. In terms of class confusion, the model's predictions align with the ground-truth labels. Furthermore, the existing confusions are reasonable. For instance, when using STEGO, confusions of the two "food" mid-level classes emerge. Overall, Fig. 8a indicates that our method either enhances or at least maintains the segmentation performance per class in terms of IoU regardless of the backbone model. Additionally, our method aids in reducing the confusion among classes.

Figure 8: **COCO-Stuff – Comparison of the segmentation performance for individual classes** in terms of IoU (in %) *(a)* and class confusion for the DINO ViT-B/8 Baseline *(b)*, PriMaPs-EM *(c)*, STEGO *(d)*, and STEGO + PriMaPs-EM *(e)*. Overall, PriMaPs-EM preserves or boosts the individual class IoU across most classes and moderately reduces class confusion.

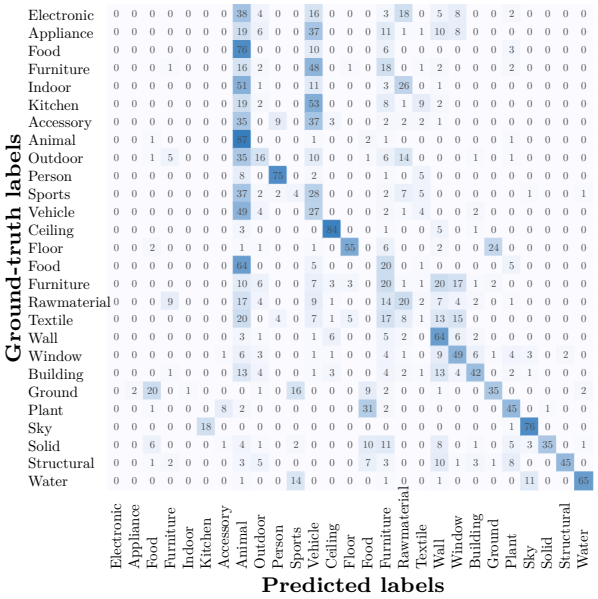
(a) Class IoUs (in %) for DINO ViT-B/8 Baseline, PriMaPs-EM *(Ours)*, STEGO, and STEGO+PriMaPs-EM *(Ours)*.

	Electronic	Appliance	Food	Furniture	Indoor	Kitchen	Accessory	Animal	Outdoor	Person	Sports	Vehicle	Ceiling	Floor	Food	Furniture	Rawmaterial	Textile	Wall	Window	Building	Ground	Plant	Sky	Solid	Structural	Water
Baseline	0.0	0.0	0.0	0.2	0.0	0.0	0.0	15.3	0.1	58.3	0.5	9.1	23.7	3.1	2.2	8.3	12.2	0.2	30.7	11.7	31.9	21.5	33.8	77.4	15.0	27.1	45.1
Ours	0.0	0.2	0.0	0.8	0.0	0.0	0.0	21.4	4.6	68.7	0.5	12.6	36.1	44.4	0.0	12.1	14.0	1.1	45.7	15.9	38.1	33.3	42.9	69.1	29.1	38.9	61.4
STEGO	0.0	0.3	0.3	13.7	0.1	0.0	0.7	74.1	0.1	61.7	10.9	40.7	36.3	30.2	38.8	22.5	15.6	11.0	36.0	2.7	51.8	44.1	50.2	82.9	19.8	37.8	58.8
STEGO+*Ours*	0.2	0.0	0.0	14.1	0.2	0.0	0.3	76.8	10.9	64.3	0.9	53.8	48.2	31.9	39.4	25.4	16.0	19.7	36.1	0.8	55.6	44.9	51.2	83.8	27.4	36.2	62.6

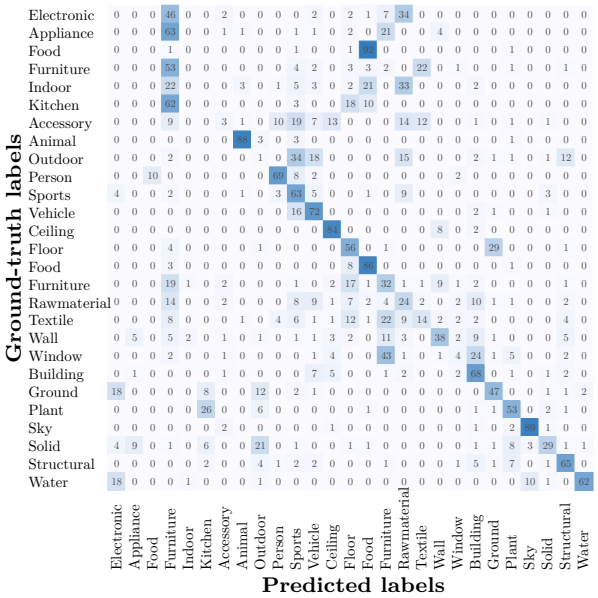
(b) Confusion matrix for the Baseline



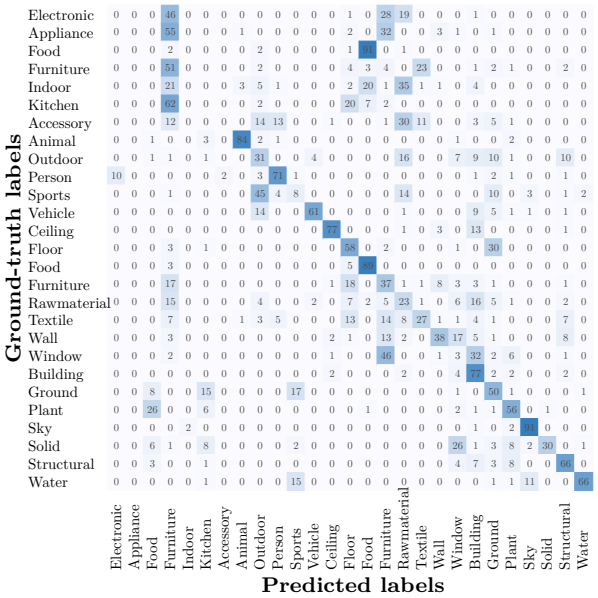
(c) Confusion matrix for PriMaPs-EM



(d) Confusion matrix for STEGO



(e) Confusion matrix for STEGO + PriMaPs-EM



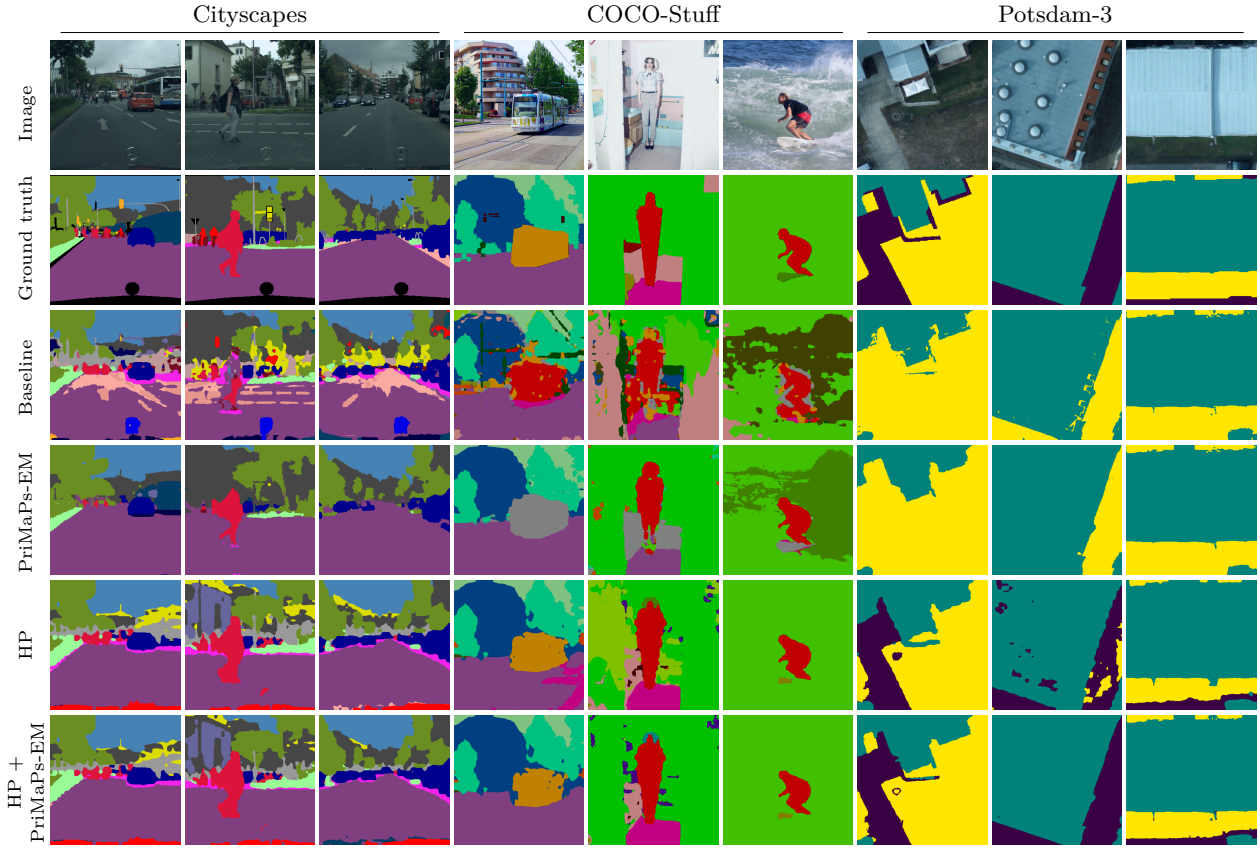


Figure 9: **Qualitative results** for the DINO ViT-S/8 Baseline, PriMaPs-EM *(Ours)*, HP (Seong et al., 2023), and HP (Seong et al., 2023)+PriMaPs-EM *(Ours)* for all three datasets. Our method produces locally more consistent segmentation results, reducing misclassification.

B.2 Qualitative Comparison to HP

Similar to the qualitative comparison in Fig. 4 in the main paper, we aim to compare PriMaPs-EM with HP (Seong et al., 2023). We present qualitative examples for the baseline, PriMaPs-EM, HP, and the combination of HP and PriMaPs-EM in Fig. 9. These qualitative examples align with the findings from the quantitative results in the main paper (*cf.* Tabs. 1 to 3). It is evident that our method produces locally more consistent and less noisy results compared to both baselines. Despite the already impressive qualitative results of HP on COCO-Stuff, our method excels in correcting misclassifications and achieving better segmentation of object boundaries. We observe one limitation, particularly with the DINO ViT-S/8 baseline, both independently and in conjunction with PriMaPs-EM for the Potsdam-3 dataset. In this case, the semantic concept of "street" is not recognized and is consequently rarely predicted.

B.3 Comparison to HP using DINOv2

Current state-of-the-art methods Hamilton et al. (2022); Seong et al. (2023) for unsupervised segmentation do not provide experiments using DINOv2 features. To be able to compare with previous methods, we train HP using DINOv2 for the comparison in Tab. 6. We strictly follow the training schedule and hyperparameters used in the original implementation for all respective datasets. HP does not generalize well to the DINOv2 features and hardly keeps up with the strong baseline. PriMaPs-EM moderately but consistently improves upon both DINOv2 and HP across all datasets and metrics using the identical PriMaPs-EM hyperparameters we use across all other experiments in this work.

Table 6: **Comparing PriMaPs-EM *(Ours)* to HP using DINOv2** for the Cityscapes (ViT-B/14), COCO-Stuff (ViT-B/14), and Potsdam-3 (ViT-S/14) datasets. We report Accuracy and mean IoU (in %) for unsupervised probing.

Method	*Cityscapes*		*COCO-Stuff*		*Potsdam-3*	
	Acc	**mIoU**	**Acc**	**mIoU**	**Acc**	**mIoU**
DINOv2 Baseline (Oquab et al., 2024)	49.5	15.3	44.5	22.9	82.4	69.9
+ HP (Seong et al., 2023)	67.9	15.9	48.9	19.8	79.4	65.7
+ PriMaPs-EM	71.6	**19.0**	46.4	**23.8**	**83.1**	**71.0**
+ HP (Seong et al., 2023) + PriMaPs-EM	**74.3**	16.6	**49.3**	20.2	79.6	66.0

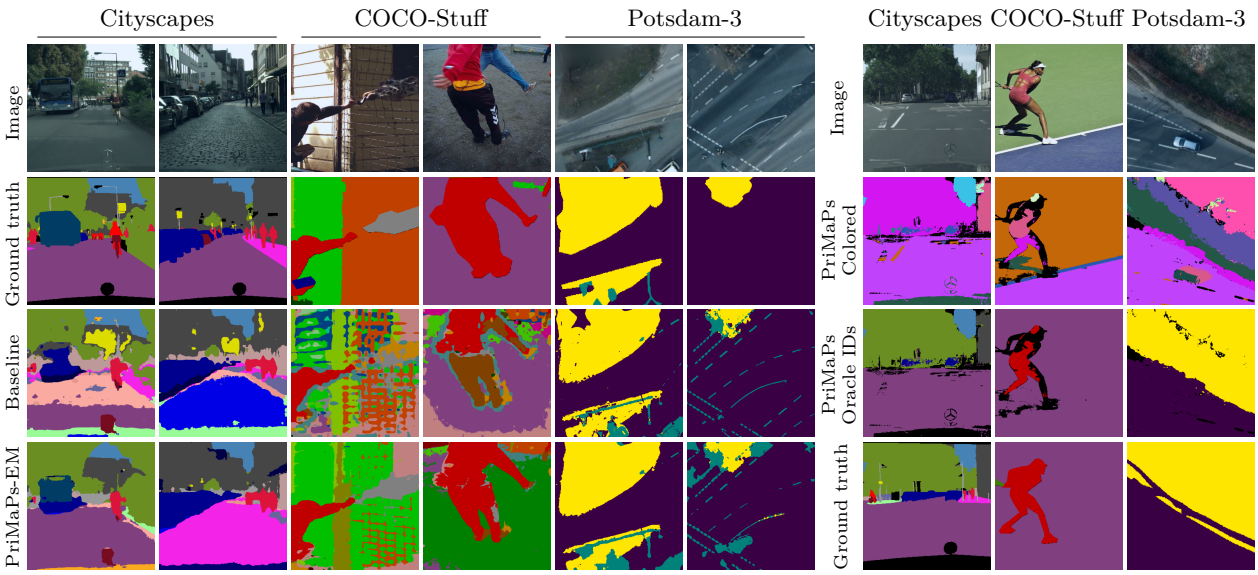


Figure 10: **Failure cases** for the PriMaPs-EM segmentation *(left)* as well as PriMaPs pseudo labels *(right)* using DINO ViT-B/8 for Cityscapes, COCO-Stuff, and Potsdam-3.

B.4 Failure Cases

Finally, we would like to discuss observed failure cases of PriMaPs-EM. Fig. 10 shows examples of failure cases occurring in the segmentation predictions as well as failure examples for PriMaPs pseudo labels. For PriMaPs-EM, we observe misclassifications, such as in Cityscapes where buses are often partially segmented as the class "car" or that cobblestone is misclassified as "sidewalk". For COCO-Stuff, shadows and structures influence the segmentation predictions and confusions also occur (see "ground" and "floor" in example two). For Potsdam-3, we observe that vehicles and road markings are sometimes erroneously attributed to the class "building" instead of "road". For PriMaPs pseudo labels, we identify two main sources of error. First, small objects in cluttered images are sometimes assigned to larger neighboring masks, a phenomenon that can be attributed to the limited backbone feature resolution. This is particularly noticeable for Cityscapes, where the center horizontal area is often detailed (*cf.* Fig. 3 "Pole", "Traffic Light", "Traffic Sign"). Second, PriMaPs sometimes oversegment images containing a large foreground object as seen in the COCO-Stuff example. Similarly, different visual appearances of the same semantic class can lead to multiple masks (Potsdam-3). Despite these observations, the simple PriMaPs provide promising mask proposals that correspond well with the ground-truth label.

C Implementation

Since all significant high-level implementation details and hyperparameters have been addressed in the main paper, this section addresses only few remaining details. Please note that we will make both the code and models publicly available upon the acceptance of this work. Our work is implemented in PyTorch (Paszke et al., 2019). We build up on the code of Ji et al. (2019), Van Gansbeke et al. (2021) and Hamilton et al. (2022).

C.1 Backbone Models

For each backbone model, we use the corresponding original implementation. Specifically, for DINO (Caron et al., 2021) and DINOv2 (Oquab et al., 2024), we utilize the PyTorch Hub implementation. In the case of STEGO (Hamilton et al., 2022) and HP (Seong et al., 2023), we integrate the respective original implementations into our framework.

C.2 Datasets

We close with some further details regarding the datasets used.

Cityscapes (Cordts et al., 2016) is an ego-centric street-scene dataset containing 5000 high-resolution images with 2048×1024 pixels. It is split into 2975 train, 500 val, and 1525 test images. Following previous work (Ji et al., 2019; Cho et al., 2021; Yin et al., 2022; Hamilton et al., 2022; Seong et al., 2023), evaluation is conducted on the 27 classes setup using the val split.

COCO-Stuff (Caesar et al., 2018) is a dataset of everyday life scenes containing 80 things and 91 stuff classes. Following previous work (Ji et al., 2019; Cho et al., 2021; Hamilton et al., 2022; Yin et al., 2022; Li et al., 2023; Seong et al., 2023), we use a reduced variant by Ji et al. (2019) containing 49629 train and 2175 test images. Hereby, all images consist of at least 75 % stuff pixels. The dataset is evaluated on the 27 classes setup.

Potsdam-3 (ISPRS) is a remote sensing dataset consisting of 8550 RGBIR satellite images with 200×200 pixels, which is split into 4545 train and 855 test images, as well as 3150 additional unlabeled images. In our experiments, the 3-label variant of Potsdam is evaluated and the additional unlabeled images are not used.