DEARLi: Decoupled Enhancement of Recognition and Localization for Semi-supervised Panoptic Segmentation

Ivan Martinović[1] Josip Šarić[2] Marin Oršić[1] Matej Kristan[2] Siniša Šegvić[1†]

[1]Faculty of Electrical Engineering and Computing
University of Zagreb
name.surname@fer.hr

[2]Faculty of Computer and Information Science
University of Ljubljana
name.surname@fri.uni-lj.si

Abstract

Pixel-level annotation is expensive and time-consuming. Semi-supervised segmentation methods address this challenge by learning models on few labeled images alongside a large corpus of unlabeled images. Although foundation models could further account for label scarcity, effective mechanisms for their exploitation remain underexplored. We address this by devising a novel semi-supervised panoptic approach fueled by two dedicated foundation models. We enhance recognition by complementing unsupervised mask-transformer consistency with zero-shot classification of CLIP features. We enhance localization by class-agnostic decoder warm-up with respect to SAM pseudo-labels. The resulting decoupled enhancement of recognition and localization (DEARLi) particularly excels in the most challenging semi-supervised scenarios with large taxonomies and limited labeled data. Moreover, DEARLi outperforms the state of the art in semi-supervised semantic segmentation by a large margin while requiring 8× less GPU memory, in spite of being trained only for the panoptic objective. We observe 29.9 PQ and 38.9 mIoU on ADE20K with only 158 labeled images. The source code is available at `github.com/helen1c/DEARLi`.

Figure 1. Unlike semi-supervised-trained state-of-the-art panoptic model [9], our recognition-enhanced method DEAR correctly detects *chair* and *bench* segments, while additional localization enhancement in DEARLi further improves masks accuracy.

1. Introduction

Panoptic segmentation assigns a semantic label to each image pixel while also distinguishing instances of object classes [31]. This is a core capability required across a wide range of applications such as autonomous driving [79], remote sensing [12], medical imaging [4], and maritime obstacle detection [85]. Strong application potential has driven considerable research efforts [7, 8, 30], which relied on vast amounts of annotated data [11, 42]. However, manual panoptic annotation is a time-consuming and tedious task, often requiring more than an hour per image [11, 56].

In related fields such as image classification [58], optical flow estimation [34], object detection [71], and semantic segmentation [48], semi-supervised learning has been extensively studied to reduce the data annotation requirements. However, only a few studies have considered semi-supervised panoptic segmentation [6, 24, 39, 49]. Importantly, these studies do not tackle the most challenging yet practical scenario, which involves very few labeled images and large class taxonomies. This setup makes labeled examples per class extremely scarce and thus requires methods with very strong generalization capabilities.

Foundation models offer a unique source of auxiliary learning signal that could be exploited to address the exposed issue. In particular, their robust representations obtained through large-scale pretraining can provide potential to improve generalization for underrepresented classes. Recently, SemiVL [22] leveraged CLIP [52] for backbone initialization and pseudo-label acquisition in semi-supervised semantics. However, dense zero-shot CLIP predictions are considerably noisy due to poor localization [70, 76].

Thus, new techniques are required to selectively pry out the skills that are embedded in foundation models and relevant for a specific performance aspect of the trained network. Effective knowledge distillation for panoptic segmentation, in particular should focus on two key aspects: (i) recognition, and (ii) localization. Distilling recognition should aid generalization across a diverse class taxonomy, while distilling localization should improve detection of both semantic and instance boundaries, thereby mitigating biases introduced by learning from limited annotations.

We address the aforementioned challenges by proposing Decoupled EnhAncement of Recognition and Localization (DEARLi) – a novel semi-supervised panoptic segmentation method driven by orthogonal contribution from two dedicated foundation models within the mask transformer framework [8]. First, we exploit a vision-language model [52] exclusively for recognition signals by ensembling its zero-shot mask-wide posteriors [70] with mask-transformer classification. Second, we leverage class-agnostic SAM [32] to generate segmentation signals. We achieve this through decoder warm-up and show that it is possible to improve a mask transformer with localization-specific pre-training. As shown in Figure 1, the recognition-enhanced model DEAR improves mask classification, while the localization enhancement in DEARLi further improves the mask segmentation.

Beyond our methodological contributions, we report the first extensive study on semi-supervised panoptic segmentation in scenarios with low annotation budgets and rich class taxonomies. The results indicate that semi-supervised DEARLi consistently outperforms supervised counterparts trained on 4× more labeled data. On ADE20K, DEAR surpasses the state of the art in semi-supervised semantic segmentation [22] by 7 mIoU points on average, despite being trained for the panoptic objective. Remarkably, our method achieves these gains using 8× less GPUs.

2. Related Work

Panoptic segmentation. Early panoptic approaches either adapt instance [21, 31] or semantic segmentation methods [5, 7]. These approaches introduce additional outputs that require expensive heuristic-based decoding. Inefficient modeling can be avoided by associating instance pixels with object queries [3] through cross-attention [62]. This incorporates a special kind of inductive bias where similar pixels of an object class should belong to the same instance. Mask Transformers (MT) [8, 9, 64, 75] apply this idea to the panoptic segmentation task. MaskFormer [8] produces a fixed number of mask embeddings together with the corresponding mask-wide class posteriors. It recovers dense mask-assignment maps by scoring high resolution features with mask embeddings. Mask2Former (M2F) [9] drives mask embeddings towards particular instances by attending queries to multiscale features through masked cross-attention. Our method builds upon Mask2Former due to reasonable priors, high accuracy and acceptable training efficiency. Our multi-stage learning pipeline (Fig. 2) extends mask transformers with orthogonal recognition and localization enhancement, which was not previously explored.

Dense prediction using vision-language models. Joint language-image representations [26, 52] led to advances in tasks such as text-to-image generation [53, 55], zero-shot classification [26, 52, 67, 80], and captioning [72]. However, nonlinear attention before the output projection hampers dense spatial inference with vision transformers [36, 37, 63, 83]. Therefore, we pair the M2F decoder with frozen ConvNeXt-CLIP [10, 43] backbone and collect per-mask embeddings from frozen features by mask pooling [17, 70, 76]. Our model architecture is similar to FC-CLIP [76] that considers supervised open-vocabulary segmentation [14, 27, 41], but does not learn on unlabeled images. FC-CLIP employs a frozen CLIP backbone with modified M2F decoders. In contrast, we propose class-agnostic pre-training of the standard M2F decoder, and show that geometric ensembling of M2F and CLIP posteriors [76] excels on underrepresented classes in the semi-supervised context.

Semi-supervised Segmentation. Only a few methods consider semi-supervised panoptics [6, 46, 49]. However, these approaches rely on a substantial amount of annotated data and outdated backbones. In contrast, we focus on low-label regimes in combination with rich class taxonomies. Our setup enables direct comparison with the related semantic segmentation works, as discussed next. These approaches leverage consistency [2, 19, 35, 45, 48], co-training [1, 51], self-training [28, 73, 77] and adversarial training [18, 47, 50, 59]. Many of these methods rely on pseudo-labels [40, 60, 65, 68, 84]. However, pseudo-label quality might hurt training due to poor accuracy or class imbalance [23, 40, 68, 84]. Our method addresses this issue through geometric ensembling [17, 20, 33, 70, 76]. Our method is closely related to methods that enforce consistency under input augmentations [58, 69, 74, 81]. We leverage the Mean Teacher framework [61] where the teacher corresponds to the exponential moving average of the student. Our teachers receive weakly perturbed images, stop the gradients and produce hard pseudo-labels [16, 25, 58, 66, 81]. SemiVL [22] learns consistency [74] with CLIP initialization and introduces a loss to maintain alignment between visual and language embeddings. In contrast, our approach avoids feature drift as the backbone remains frozen. Compared to SemiVL, our method performs zero-shot classification on mask-pooled convolutional CLIP features instead of individual ViT tokens, which reduces noise in pseudo-labels. In addition, SemiVL's decoder builds upon per-pixel vision-language similarities, which limits its ability to capture weak correlations. Finally, our approach supports both panoptic and semantic segmentation.

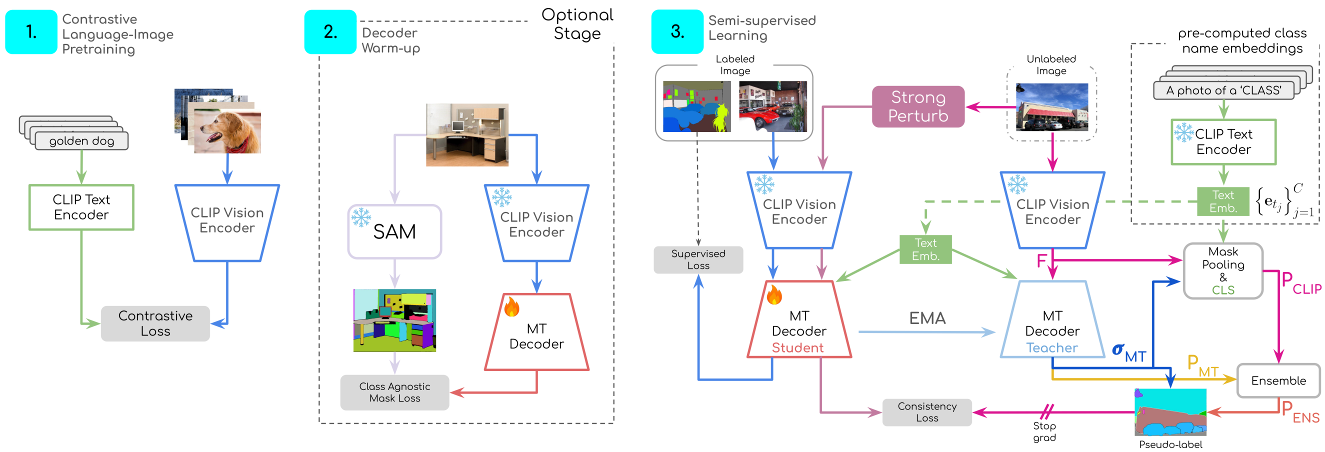

Figure 2. Overview of our three-stage semi-supervised learning pipeline. The first stage corresponds to large-scale contrastive language-image pre-training (CLIP [10, 52]). The second stage enhances the localization by class-agnostic mask transformer (MT) decoder warm-up. The third stage trains the decoder on labeled and unlabeled images with Mean Teacher consistency. We enhance the teacher recognition by ensembling the mask-wide posterior of the mask transformer with zero-shot classification of mask-pooled CLIP features.

3. Method

Semi-supervised learning (SSL) involves a small amount of labeled data $\mathcal{X}^l = \left\{\left(x_i^l, y_i^l\right)\right\}_{i=1}^{N_l}$ and a large amount of unlabeled data $\mathcal{X}^u = \{x_i^u\}_{i=1}^{N_u}$, where $N_l \ll N_u$. The goal is to utilize both subsets to achieve better generalization performance. These algorithms typically optimize a compound loss on labeled and unlabeled samples:

$$\mathcal{L}^{\text{SSL}} = \mathcal{L}^{\text{SUP}} + \mathcal{L}^{\text{UNL}}. \tag{1}$$

The first component corresponds to the standard supervised loss $\mathcal{L}_i^{\text{SUP}} = \mathcal{L}(o_i, y_i^{GT})$, where y_i^{GT} denotes the ground truth and $o_i = h_{\theta_{stud}}\left(x_i^l\right)$ represents the model prediction computed in the corresponding labeled example x_i^l. We express the contribution of the unlabeled data with the consistency loss $\mathcal{L}_i^{\text{UNL}} = \mathcal{L}(o_i^u, \hat{y}_i^u)$ [69]. The model output $o_i^u = h_{\theta_{stud}}\left(\mathcal{S}(x_i^u)\right)$ is computed with respect to strongly perturbed unlabeled example $\mathcal{S}(x_i^u)$, while the pseudo-label $\hat{y}_i^u = h_{\theta_{teach}}\left(\mathcal{W}(x_i^u)\right)$ is derived from the same, but weakly perturbed example $\mathcal{W}(x_i^u)$. The teacher $h_{\theta_{teach}}$ and the student $h_{\theta_{stud}}$ follow the same architecture and therefore can be plugged into the Mean Teacher framework [61]. Each iteration t sets θ_{teach} to the exponential moving average (EMA) of θ_{stud} with decay rate γ:

$$\theta_{teach}^t = \gamma\theta_{teach}^{t-1} + (1-\gamma)\,\theta_{stud}^{t-1}. \tag{2}$$

Figure 2 provides an overview of our training pipeline. The first stage corresponds to large-scale contrastive language-image pre-training (CLIP) [52]. The resulting model serves as a backbone feature extractor of our panoptic models. These models build upon mask transformers [9, 38, 64, 75] in order to enable elegant integration of separate recognition and segmentation foundation models, as explained in 3.1. The second stage enhances the localization by warming-up the mask transformer decoder on class-agnostic regions, which we discuss in 3.3. The third stage trains the mask transformer decoder on the target dataset according to the semi-supervised objective (Eq. 1). Here, we enhance the recognition by ensembling the standard teacher's mask-wide posteriors with zero-shot classification of mask-pooled CLIP features as discussed in Sec. 3.2.

3.1. Panoptic Mask Transformers

Mask transformers [9, 38, 64, 75] start from N learnable queries, attend them to translation equivariant features, and predict mask embeddings of *thing* and *stuff* classes. Mask embeddings are then projected onto mask-wide logits $\mathbf{P} \in \mathbb{R}^{N\times(C+1)}$ and used to score high resolution features into sigmoid-activated localization maps $\boldsymbol{\sigma} \in \mathbb{R}^{N\times H\times W}$. The C+1-th *no-object* class allows to discard excess queries since N is usually larger than the number of segments in the image. Building on the set prediction framework [3], mask transformers establish bipartite matching $\mathcal{M}$ between predicted masks and ground truth segments in order to compute the loss. The compound loss consists of recognition and localization terms:

$$\mathcal{L} = \sum_i^N \mathcal{L}_{cls}(\mathbf{P}_i, y_{\mathcal{M}(i)}^{GT}) + \sum_{y_{\mathcal{M}(i)}^{GT} \neq C+1} \mathcal{L}_{loc}(\boldsymbol{\sigma}_i, \boldsymbol{\sigma}_{\mathcal{M}(i)}^{GT}). \tag{3}$$

Here $\mathcal{L}_{cls}$ denotes segment-wide recognition loss expressed with standard cross-entropy, while $\mathcal{L}_{loc}$ corresponds to per-pixel localization loss expressed as a combination of the dice loss and binary cross-entropy.

3.2. Vision-Language Enhanced Recognition

Mask transformers can easily overfit to limited labeled data. This leads to biased pseudo-labels in semi-supervised learning. To address this issue, we integrate mask transformer with a CLIP foundation model, which has less pronounced biases due to large-scale pretraining.

We begin by extracting image features with a frozen CLIP backbone. While parameter freezing limits the learning capacity, it offers several benefits. First, it preserves zero-shot mask classification ability since the model embeds images into the joint vision-text feature space [52]. Second, it decreases the risk of overfitting and bias absorption due to insufficient labeled training data [36, 63, 83]. Third, it reduces the training footprint since backpropagation through the CLIP backbone becomes unnecessary.

Next, we fix the final layer weights of the mask classifier to the class embeddings from the CLIP text encoder. Consequently, the mask decoder must learn to map mask embeddings into the language space for accurate classification. Built on Mask2Former [9], we call this model M2F-Lang.

Finally, rather than relying solely on the mask transformer for pseudolabel generation, we ensemble its classification probabilities with those from zero-shot evaluated CLIP. Since CLIP provides only image-level recognition, we first describe the mask pooling [17, 70, 76] operation that enables zero-shot mask-level classification. Given dense CLIP features $\mathbf{F} \in \mathbb{R}^{H' \times W' \times D}$ and a binary mask $\mathbf{M}_i \in \{0,1\}^{H \times W}$, mask pooling recovers mask-aggregated visual embedding $\mathbf{e}_{v_i} \in \mathbb{R}^D$ as follows:

$$\mathbf{e}_{v_i} = \mathcal{MP}\left(\mathbf{F}, \mathbf{M}_i\right) = \frac{\sum_{r,c}^{HW} \mathbf{F}[r,c,:] \cdot \mathbf{M}_i[r,c]}{\sum_{r,c}^{HW} \mathbf{M}_i[r,c]}. \quad (4)$$

We downsample the masks to match feature resolution. The visual embedding $\mathbf{e}_{v_i}$ is then compared against a set of class embeddings $\left\{\mathbf{e}_{t_j}\right\}_{j=1}^{C}$ obtained by applying the CLIP text encoder $\mathcal{T}$ to class names. This comparison yields a class probability distribution, computed as the softmax of cosine similarities scaled by a temperature parameter τ:

$$\mathbf{p}^i_{\text{CLIP}} = \text{softmax}([\mathbf{e}_{v_i}^T \mathbf{e}_{t_1}, \mathbf{e}_{v_i}^T \mathbf{e}_{t_2}, ..., \mathbf{e}_{v_i}^T \mathbf{e}_{t_C}], \tau). \quad (5)$$

The next paragraph explains how CLIP mask-pooling can enhance pseudo-labels for semi-supervised learning.

Given a weakly augmented unlabeled image $\mathcal{W}\left(x^u\right)$, the teacher network first generates N initial mask candidates. After removing those classified as *no object*, a set of N' masks remain, defined with sigmoid-activated localization maps $\boldsymbol{\sigma}_{\text{MT}} \in \mathbb{R}^{N' \times H \times W}$ and class probabilities $\mathbf{P}_{\text{MT}} \in \mathbb{R}^{N' \times C}$. We then calculate a binary mask $\mathbf{M}_i \in \{0,1\}^{H \times W}$ for each of the N' masks by thresholding the corresponding sigmoid mask: $\mathbf{M}_i = [\![\boldsymbol{\sigma}_{\text{MT}i} \geq 0.5]\!]$. Next, we recover the cached CLIP features $\mathbf{F} \in \mathbb{R}^{H' \times W' \times D}$ from the frozen backbone and retrieve zero-shot class distributions $\mathbf{P}_{\text{CLIP}} \in \mathbb{R}^{N' \times C}$ via mask pooling over all masks. Finally, we determine the ensembled mask-wide posteriors as a weighted geometric mean of the mask transformer posteriors and the zero-shot posteriors (5) [17, 70, 76]:

$$\mathbf{P}_{\text{ENS}} = \left(\mathbf{P}_{\text{MT}}\right)^{\alpha} \odot \left(\mathbf{P}_{\text{CLIP}}\right)^{1-\alpha}. \quad (6)$$

These ensembled probabilities are then fed to the standard panoptic inference [8] to recover hard pseudo-labels for consistency training. Such pseudo-label acquisition exploits only the recognition signal from CLIP as boundaries are retrieved from the mask-transformer, which was originally intended. Although the student learns from enhanced pseudo-labels, we observe slight improvements when including the ensembling during inference (*cf*. Tab. 14 in suppl.).

Validating CLIP zero-shot mask classification. As a proof of concept, we test the CLIP zero-shot recognition capabilities in the panoptic context. In particular, we measure the panoptic quality of a model with CLIP features $\mathbf{F} \in \mathbb{R}^{H' \times W' \times D}$ and the oracle mask proposal generator. We extract dense CLIP features with MaskCLIP [83] for ViT [15, 52] and simply remove global average pooling for ConvNeXt [10]. The oracle generator retrieves a binary mask candidate $\mathbf{M} \in \{0,1\}^{H \times W}$ for each ground truth segment. We recover the per-mask class distributions with mask pooling (Eqs. 4 and 5).

Figure 3 presents the resulting panoptic quality on ADE20K and COCO-Panoptic. We observe that zero-shot classification of mask-pooled CLIP features delivers attractive performance. This shows that mask transformer can benefit from the ensembling. Just as important, ConvNeXt models significantly outperform ViT on both datasets. This motivates us to reconsider the ViT backbone used in the previous state-of-the-art for semi-supervised semantics [22]. Inspired by these findings, we select the frozen OpenCLIP ConvNeXt [10] pretrained on LAION-2B [57] as the backbone and include zero-shot classification of mask-pooled CLIP features in our pseudo-label generation procedure.

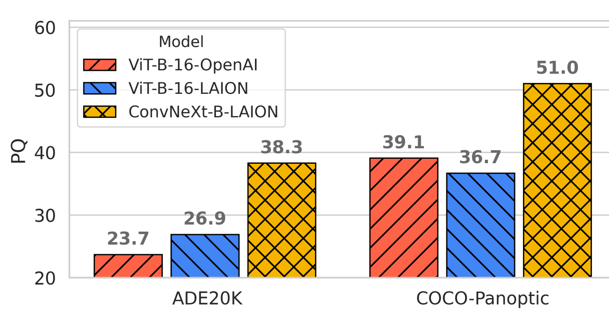

Figure 3. Zero-shot panoptic quality with ground-truth masks and mask-pooled CLIP features. Each mask is classified with respect to textual CLIP embeddings of class name descriptions.

3.3. Class-agnostic Mask-Decoder Warm-up

The proposed ensembling with zero-shot predictions assumes adequate mask candidates with accurate boundaries. Failure of this assumption would devalue the benefits of the proposed ensembling. This risk is especially pertinent to semi-supervised setups where the model may struggle to learn precise object boundaries due to limited labeled data.

We address this challenge by distilling objectness from the Segment Anything Model (SAM) [32]. Given an unlabeled image, SAM can produce a set of class-agnostic binary masks $\{\mathbf{M}_i\}_{i=1}^{N_{\text{SAM}}}$ corresponding to different regions in the image. A naive approach to obtain panoptic pseudo-labels classifies each SAM mask $\mathbf{M}_i$ with CLIP using the mask pooling (Eqs. 4 and 5). However, this approach yields only 8.8 PQ on ADE20K. This poor performance is a result of a significant granularity mismatch between SAM predictions and the target dataset, as illustrated in Fig. 4.

We propose a simple yet effective solution to leverage SAM's rich objectness knowledge while simultaneously addressing the granularity mismatch issue. The key idea is to precondition the mask decoder prior to the main semi-supervised stage. Specifically, we first generate class-agnostic pseudo-labels using SAM for both labeled images $\mathcal{X}^l$ and unlabeled images $\mathcal{X}^u$. These class-agnostic pseudo-labels are then used to warm up the mask decoder, as illustrated in the middle pane of Fig. 2. During this warm-up phase, the optimization focuses solely on the localization loss term $\mathcal{L}_{loc}$ (*cf*. Eq. 3), leaving classification and granularity refinement to be learned later from the available labeled subset. The proposed Decoder Warm-up (DeWa) stage enhances both recall and boundary alignment of the mask decoder, as demonstrated in our experiments.

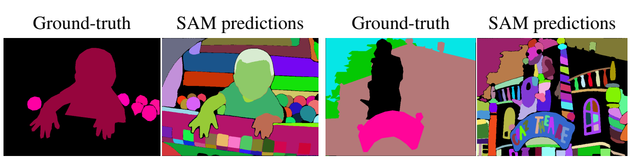

Figure 4. Illustration of granularity mismatch between ADE20K GT labels and class-agnostic predictions generated by SAM [32].

4. Experiments

Datasets. We conduct experiments on: ADE20K, COCO-Panoptic, and COCO-Objects. The ADE20K dataset [82] comprises 20 210 training and 2000 validation images with 150 semantic classes. COCO-Panoptic [42] includes 118 k training and 5 k validation images, with 80 *thing* and 53 *stuff* classes. COCO-Objects [42] corresponds to the same dataset with all *stuff* classes remapped to the *background* class. We consider this dataset to ensure fair comparison with semi-supervised semantic segmentation methods.

Architecture. Our experiments build upon ConvNeXt-B (CN-B) [43] and assume CLIP [10, 52] pre-training on LAION-2B (L2B) [57]. We complement this backbone with the M2F [9] decoder. Models with language-based recognition (M2F-Lang) produce classification logits using dot product between mask-wide class embeddings and precomputed language embeddings. Effectively, frozen language embeddings operate as a hand-crafted linear projection.

Training. Our semi-supervised approach uses Mean Teacher [61] across weakly and strongly perturbed unlabeled images (Sec. 3). Weak perturbations are random scaling (0.1 to 2.0), cropping and horizontal flipping. Strong perturbations further include color jittering, Gaussian blur, grayscaling and CutMix [78]. We train for 80 000 iterations in all semi-supervised experiments. Training batches consist of 8 labeled and 8 unlabeled images. The student and teacher networks share the same architecture. Teacher network parameters are updated using EMA with $\gamma = 0.999$.

We use identical dataset partitions as in the prior work [22, 74]. We train on crop sizes 640×640 and 512×512 for ADE and COCO, respectively. We use the AdamW [29, 44] optimizer with weight decay 0.05 and learning rate 0.0001. Unless otherwise specified, we freeze the backbone and train only the M2F decoder. We set geometric ensembling factor $\alpha = 0.6$ across all experiments. For language-based recognition, we adopt prompt templates [17, 20, 70]. In M2F-Lang experiments, we do not reject segments with low max-softmax [76]. We warm-up the decoder (DeWa) with batch size 16 for 80 k and 160 k iterations on ADE and COCO, respectively. Class-agnostic masks are generated with the ViT-H SAM [32] and default hyperparameters. All experiments with a frozen backbone are conducted on a single A100-40 GB GPU.

Evaluation. Most of our experiments report panoptic quality (PQ), averaged across all dataset classes [31]. We compare with previous semantic segmentation approaches in terms of mean intersection over union (mIoU). We report performance of the checkpoint from the final training iteration on the ADE20K/COCO validation set.

4.1. Panoptic Segmentation Results and Ablations

Table 1 outlines the development path of our method. It presents performance increments and component ablations from our mask-recognition baseline (M2F) over the CLIP enhanced DEAR, to the CLIP and SAM enhanced DEARLi. All experiments report PQ performance on ADE20K val and train on standard semi-supervised partitions [74]. We evaluate how CLIP initialization, language-based recognition, per-class probability geometric ensembling ($\mathbf{P}_{\text{ENS}}$) and decoder warm-up (DeWa) contribute to our method.

We observe that semi-supervised learning (SSL✓) consistently outperforms fully supervised counterparts. This provides a sanity check for further experiments. Gains

Method	SSL	Backbone	1/128 (158)			1/64 (316)			1/32 (632)			1/16 (1263)			1/8 (2526)		
			mPQ	mSQ	mRQ	**mPQ**	mSQ	mRQ	**mPQ**	mSQ	mRQ	**mPQ**	mSQ	mRQ	**mPQ**	mSQ	mRQ
M2F	–	❄ CN-B-IN1k	7.4	32.7	9.3	13.3	52.2	16.7	17.8	62.7	21.7	22.1	65.1	27.0	25.5	69.3	30.7
	–	🔥 CN-B-IN1k	9.3	38.6	11.7	15.0	55.2	18.6	20.5	62.9	25.2	25.0	68.9	30.5	29.9	73.2	35.9
	✓	🔥 CN-B-IN1k	16.4	45.7	20.4	22.8	60.4	27.9	27.8	70.2	34.0	30.1	71.4	36.6	33.6	75.1	40.3
M2F	–	❄ CN-B-L2B	10.0	38.7	12.4	16.4	56.2	20.1	23.4	65.4	28.5	28.7	69.5	34.9	34.2	73.7	40.9
	–	🔥 CN-B-L2B	12.9	42.8	16.1	19.3	59.1	23.9	26.0	69.1	31.7	31.1	72.0	37.7	36.3	75.6	43.7
	✓	🔥 CN-B-L2B	19.4	46.2	24.0	28.4	63.7	34.6	33.2	71.3	40.2	36.8	75.9	44.3	40.2	78.8	48.2
M2F-Lang	–	❄ CN-B-L2B	11.7	43.7	14.7	19.1	61.4	23.9	26.3	70.2	32.1	31.6	72.8	38.3	36.6	76.7	44.0
↳ + $\mathbf{P}_{\text{ENS}}$ (eval only)	–		18.3	63.7	23.3	23.9	69.5	30.0	29.0	73.3	35.4	33.6	75.6	40.8	37.8	77.6	45.7
M2F-Lang	✓		18.3	55.0	22.6	26.4	68.0	32.3	32.5	76.1	39.4	35.5	74.5	42.8	39.2	77.2	47.2
↳ + $\mathbf{P}_{\text{ENS}}$ (eval only)	✓		21.6	62.6	26.7	30.2	73.9	37.0	33.9	76.2	41.1	36.7	77.6	44.2	40.2	77.2	48.4
↳ + $\mathbf{P}_{\text{ENS}}$ (**DEAR**)	✓		27.7	70.3	34.4	32.3	74.0	39.6	34.8	75.4	42.1	38.3	79.3	46.2	40.6	80.7	48.7
↳ + DeWa (**DEARLi**)	✓		**29.9**	74.0	36.6	**34.6**	75.4	41.2	**36.3**	77.1	43.5	**39.2**	80.3	47.1	**41.6**	80.6	49.5

Table 1. Impact of our contributions to panoptic performance on ADE20K. Top two sections start from the supervised baseline and show improvements due to backbone fine-tuning (❄→ 🔥), semi-supervised learning with Mean Teacher consistency (SSL) and LAION-2B pre-training (L2B). Bottom section involves language-based recognition (M2F-Lang) and shows improvements due to SSL, geometric ensembling with CLIP during inference (eval-only) and pseduolabels generation ($\mathbf{P}_{\text{ENS}}$), and class-agnostic Decoder Warm-up (DeWa).

from semi-supervised learning are more pronounced in low-label data regimes. Comparison between the first two sections in Tab. 1 reveals that LAION-2B CLIP initialization consistently surpasses ImageNet1k [13] initialization (from 3 p.p. PQ on 1/128 to 6.6 p.p. PQ on 1/8). Backbone fine-tuning improves performance in supervised experiments across the first two sections. However, this roughly doubles the training memory requirements. More importantly, fine-tuning the backbone introduces vision-language misalignment that degrades performance in low-label regimes (preliminary experiments show -4.8 p.p. PQ on 1/128). Hence, all experiments from the last section freeze the backbone.

The last section focuses on learning with fixed language embeddings (M2F-Lang). Here, mask-wide logits correspond to cosine similarity between retrieved mask-wide embeddings and precomputed language embeddings. Recall that mask-wide class embeddings can be recovered either from the M2F decoder or through mask pooling of CLIP features. Ensembling the corresponding two posteriors leads to consistent and substantial improvement with respect to the decoder-only posterior (M2F-Lang). M2F-Lang + $\mathbf{P}_{\text{ENS}}$ (eval only) retains some improvement even when ensembling posteriors exclusively during inference. We observe that semi-supervised learning on unlabeled images leads to similar or greater improvements as mask-wide posterior ensembling. Importantly, incorporating geometric ensembling into pseudo-labels generation provides additional significant gains, as the student benefits from a stronger learning signal. This leads us to our method, DEAR, which consistently improves upon evaluation-only ensembling across all partitions. DEAR performs especially well in the most challenging regime with only 158 labeled images. Finally, the inclusion of class-agnostic decoder warm-up (DEARLi) further improves the performance across all data partitions. This supports our hypothesis that incorporating class-agnostic pre-training with SAM pseudo-labels can improve final panoptic quality, despite the granularity mismatch between SAM pseudo-labels and ADE20K taxonomy (*cf*. Fig. 4). For more comprehensive ablations, see supplement (Tab. 14) .

Table 2 shows similar findings on COCO-Panoptic. Posterior ensembling for pseudo-label generation (DEAR) consistently outperforms baseline across all partitions. Moreover, SAM distillation within the decoder warm-up (DEARLi) yields further gains, with a notable +4.1 p.p. PQ on the most challenging setup with only 232 labeled images. Note that DEARLi surpasses its supervised counterpart (M2F-Lang+$\mathbf{P}_{\text{ENS}}$) while using 4× less labeled data.

Method	SSL	1/512 (232)	1/256 (463)	1/128 (925)	1/64 (1849)	1/32 (3697)
M2F-Lang	–	15.0	26.0	33.2	37.8	41.1
↳ + $\mathbf{P}_{\text{ENS}}$ (eval only)	–	22.9	30.7	36.3	39.5	42.5
M2F-Lang	✓	27.6	34.7	38.9	42.0	44.2
↳ + $\mathbf{P}_{\text{ENS}}$ (**DEAR**)	✓	34.7	38.6	40.8	43.0	44.8
↳ + DeWa (**DEARLi**)	✓	**38.8**	**41.3**	**43.1**	**44.5**	**46.4**

Table 2. PQ performance on **COCO-Panoptic**. All experiments leverage frozen ❄ CN-B-L2B backbone. Top section shows the supervised model and improvement due to inference only ensembling. Bottom section starts from semi-supervised learning with decoder only posteriors, and presents improvements due to synergy of SSL and $\mathbf{P}_{\text{ENS}}$ (DEAR) and decoder warm-up (DEARLi).

4.2. Comparison with the State of the Art

To the best of our knowledge, our method is the first to address semi-supervised panoptic segmentation with standard partitioning [74] of the common segmentation datasets with rich taxonomies. Consequently, we cannot present a direct comparison with previous semi-supervised panoptic approaches. Instead, we compare DEAR with the state of the art in semi-supervised semantic segmentation using the same labeled data partitions. In order to measure semantic segmentation performance, we evaluate our panoptic models with semantic inference, without retraining [8]. We consider this comparison relevant and fair since panoptic M2F models typically underperform with respect to native semantic segmentation models on that particular task [9]. For an exact comparison, see Tab. 12 in the supplement.

We are particularly interested in the comparison with SemiVL [22], since it also leverages CLIP. Tables 3 and 5 indicate that DEAR beats all previous methods by a large margin across all partitions on ADE20K and COCO-Objects. We additionally include DEARLi in these tables even though comparison with SemiVL might be unfair, since DEARLi distills knowledge from SAM. Nevertheless, it may prove as a useful baseline for future semi-supervised approaches with class-agnostic pre-training.

We observe that our baseline M2F+SSL with ImageNet1k initialization outperforms all previous baselines with the same pre-training, while even exceeding SemiVL in some assays. This suggests that M2F+SSL is a strong baseline and strengthens the value of our contributions. SemiVL with a convolutional backbone performs roughly the same as SemiVL with a transformer backbone (*cf*. Tab. 3) or even worse (*cf*. Tab. 5). This suggests that the observed advantages of the convolutional backbone (*cf*. Fig. 3) are likely due to direct segment prediction and mask pooling [9, 70] being present in our architecture. The supplement contains details of SemiVL [22] training atop CN-B-L2B (Appendix B), as well as semantic segmentation ablations of DEARLi (Tab. 13).

Method	Net	1/128 (158)	1/64 (316)	1/32 (632)	1/16 (1263)	1/8 (2526)
CutMix [16] [BMVC'20]	R101	–	–	26.2	29.8	35.6
AEL [23] [NeurIPS'21]	R101	–	–	28.4	33.2	38.0
UniMatch [74] [CVPR'23]	R101	15.6	21.6	28.1	31.5	34.6
UniMatch [22, 74] [CVPR'23]	ViT-B/16	18.4	25.3	31.2	34.4	38.0
M2F+SSL (*cf*. Tab. 1)	CN-B-IN1k	19.9	29.4	34.0	37.4	41.2
SemiVL [22] [ECCV'24]	ViT-B/16	28.1	33.7	35.1	37.2	39.4
SemiVL† [22] [ECCV'24]	CN-B-L2B	30.2	33.3	36.3	37.7	40.7
DEAR	CN-B-L2B	36.5 (+8.4)	40.5 (+6.8)	42.8 (+7.7)	45.8 (+8.6)	47.5 (+8.1)
DEARLi	CN-B-L2B	38.9 (+10.8)	42.0 (+8.3)	44.3 (+9.2)	45.0 (+7.8)	48.1 (+8.7)

Table 3. Comparison with the state of the art in semi-supervised semantic segmentation (mIoU) on **ADE20K**. † indicates our experiments with public source code. Underline denotes CLIP-WiT [52] initialization. Improvements over SemiVL-ViT-B/16 are in green. Gold, silver and bronze denote the best results.

Method	1/512 (232)	1/256 (463)	1/128 (925)	1/64 (1849)	1/32 (3697)
❄ DEAR	37.9	41.0	42.9	45.0	46.5
❄ DEARLi	42.0	43.7	45.4	47.6	48.7

Table 4. Semi-supervised PQ performance on **COCO-Objects**.

Method	Net	1/512 (232)	1/256 (463)	1/128 (925)	1/64 (1849)	1/32 (3697)
PseudoSeg [81, 84] [ICLR'21]	XC-65	29.8	37.1	39.1	41.8	43.6
PC^{2}Seg [81] [ICCV'21]	XC-65	29.9	37.5	40.1	43.7	46.1
CISC-R [68] [TPAMI'23]	XC-65	32.1	40.2	42.2	–	–
UniMatch [74] [CVPR'23]	XC-65	31.9	38.9	44.4	48.2	49.8
UniMatch [22, 74] [CVPR'23]	ViT-B/16	36.6	44.1	49.1	53.5	55.0
LogicDiag [40] [ICCV'23]	XC-65	33.1	40.3	45.4	48.8	50.5
AllSpark [65] [CVPR'24]	MiT-B5	34.1	41.7	45.5	49.6	–
S4Former [25] [CVPR'24]	DeiT-B	35.2	43.1	46.9	–	–
M2F+SSL	CN-B-IN1k	38.6	46.0	50.8	54.6	55.7
SemiVL [22] [ECCV'24]	ViT-B/16	50.1	52.8	53.6	55.4	56.5
SemiVL† [22] [ECCV'24]	CN-B-L2B	47.6	49.1	50.1	52.6	52.9
DEAR	CN-B-L2B	52.9 (+2.8)	53.9 (+1.1)	56.2 (+2.6)	58.7 (+3.3)	59.3 (+2.8)
DEARLi	CN-B-L2B	54.6 (+4.5)	55.1 (+2.3)	57.0 (+3.4)	59.1 (+3.7)	60.2 (+3.7)

Table 5. Comparison with the state of the art in semi-supervised semantic segmentation (mIoU) on **COCO-Objects**. † indicates our experiments with public source code. Underline denotes CLIP-WiT [52] initialization.

Comparison on COCO-Objects. Many previous works in semi-supervised semantic segmentation report experiments on the COCO-Objects dataset, which includes 80 *thing* classes and one *background* class. We enable training of our panoptic models on these 81 classes by remapping all *stuff* classes from COCO-Panoptic to the *background* class. Table 4 shows that class-agnostic decoder warm-up consistently enhances panoptic performance as in previous setups. Moreover, Table 5 shows that DEAR again consistently outperforms current state of the art [22] across all partitions, while DEARLi confirms the advantage from Table 4. A qualitative comparison with state-of-the-art methods on several examples is provided in the supplement.

4.3. Additional Ablations and Analysis

Decoder warm-up. Table 6 validates the performance of our method with different decoder warm-up procedures. We keep the backbone frozen. We start with randomly initialized decoder, as used in DEAR. Row 2 shows that decoder initialization with supervised pre-training (M2F-Lang from Tab. 1) actually decreases the performance, which suggests the presence of overfitting. Row 3 trains on random 110k images from SA1B, the dataset that was originally used to train SAM [32]. This experiment follows hyperparameters from DEARLi (row 4). Comparable performance of the last two rows suggests that DEAR benefits from de-

Method	1/128 (158)	1/64 (316)	1/32 (632)	1/16 (1263)	1/8 (2526)
Random init (**DEAR**)	27.7	32.3	34.8	38.3	40.6
Labeled-only init	26.7	31.7	34.8	38.0	40.4
SA1B (1%) - pretraining	**30.2**	34.0	**36.3**	38.7	**41.8**
SAM pseudo-labels (**DEARLi**)	29.9	**34.6**	**36.3**	**39.0**	41.6

Table 6. Validation of decoder warm-up procedures on ADE20K.

coder warm-up, even in the presence of domain shift from the SA1B. We highlight the last section improvement as a valuable contribution with plenty of applications.

Geometric ensembling ablation. Table 7 validates pseudolabel generation with $\mathbf{P}_{\mathrm{MT}}$, $\mathbf{P}_{\mathrm{CLIP}}$, or $\mathbf{P}_{\mathrm{ENS}}$ on ADE20K. Here, we use DEAR with standard M2F inference in the student (w/o ensembling). We observe consistent performance improvements with $\mathbf{P}_{\mathrm{ENS}}$, which justifies proposed mechanism of knowledge distillation from CLIP.

Method		1/128	1/64	1/32	1/16	1/8
$\mathbf{P}_{\mathrm{MT}}$	(teacher only)	18.3	26.4	32.5	35.5	39.2
$\mathbf{P}_{\mathrm{CLIP}}$	(teacher only)	24.1	28.3	32.5	36.1	39.2
$\mathbf{P}_{\mathrm{ENS}}$	(Eq. 6, teacher only)	**27.3**	**31.5**	**34.4**	**38.0**	**40.3**

Table 7. Ablation of posterior ensembling on ADE20K (PQ).

Gains on underrepresented classes. Figure 5 compares M2F+SSL (*cf*. Tab. 1, 6th row) with DEARLi when trained on ADE 1/128 and ADE 1/64 semi-supervised setups. We group classes by pixel ratio, with 30 classes per group from least to most frequent. Notably, 20 % of classes occupy over 80 % of labeled pixels in both setups, reflecting a long-tailed distribution. We observe substantial gains of DEARLi in underrepresented classes and smaller gains in frequent classes. As expected, this difference decreases in partitions containing more labeled data.

Increasing the backbone. Figure 6 validates the panoptic quality of DEAR and DEARLi on ADE and COCO-Panoptic when increasing the frozen backbone size from ConvNeXt-B to ConvNeXt-L (*i.e.* 88M $\rightarrow$ 197M). The larger backbone consistently improves results. DEARLi surpasses DEAR in all configurations. DEARLi particularly excels in very low-labeled regimes, achieving performance improvements comparable to doubling the backbone capacity. The supplement includes exact measurements.

Validating α. Table 8 evaluates panoptic quality of DEAR on ADE20K as the geometric ensembling factor α varies. We find that $\alpha = 0.6$ offers the best balance across data splits, with higher values improving performance on larger partitions (e.g., 1/8). This aligns with expectations, as higher α increases reliance on learned classification, which tends to be more accurate with more labeled data.

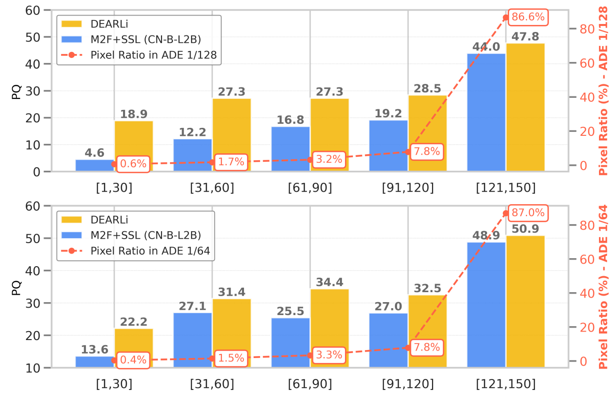

Figure 5. PQ performance on ADE 1/128 (top) and ADE 1/64 (bottom) for five groups of classes ranked according to per-pixel label incidence. DEARLi improves upon M2F+SSL in all groups, but most significantly on underrepresented classes.

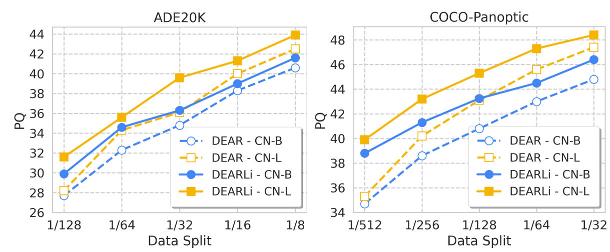

Figure 6. PQ performance of DEAR and DEARLi with two different backbones on ADE20K (left) and COCO-Panoptic (right).

	1/128 (158)	1/64 (316)	1/32 (632)	1/16 (1263)	1/8 (2526)
$\alpha = 0.5$	26.70	31.22	34.60	36.88	39.66
$\alpha = 0.6$	**27.67**	**32.27**	34.77	**38.27**	40.62
$\alpha = 0.7$	24.35	31.05	**34.88**	37.80	**41.42**

Table 8. PQ performance of DEAR on ADE20K for varying α.

5. Conclusion

We have presented a foundation-model-powered method for semi-supervised panoptic segmentation. Our method assumes decoupled recognition and localization heads of a mask transformer baseline and enhance them separately. Recognition is enhanced by ensembling the standard mask-wide posteriors with zero-shot classification of mask-pooled CLIP features. Localization is improved through class-agnostic decoder warm-up towards SAM pseudo-labels. Our panoptic models establish strong semi-supervised baselines on ADE20K and COCO, surpassing supervised counterparts trained with 4× more labeled data. They also achieve state-of-the-art performance in semi-supervised semantic segmentation despite optimizing a panoptic objective, while requiring 8× fewer GPUs.

Acknowledgments

This work has been supported by Croatian Recovery and Resilience Fund - NextGenerationEU (grant C1.4 R5-I2.01.0001), Croatian Science Foundation (grant IP-2020-02-5851 ADEPT), Advanced computing service provided by the University of Zagreb University Computing Centre - SRCE, Slovenian research agency research program P2-0214, and European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Postdoctoral Fellowship Programme, SMASH co-funded under the grant agreement No. 101081355. The SMASH project is co-funded by the Republic of Slovenia and the European Union from the European Regional Development Fund. In memory of dear mentor, colleague and friend, Siniša Šegvić.

References

[1] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, 1998. 2

[2] Gerda Bortsova, Florian Dubost, Laurens Hogeweg, Ioannis Katramados, and Marleen De Bruijne. Semi-supervised medical image segmentation via learning consistency under transformations. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*, pages 810–818. Springer, 2019. 2

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2, 3

[4] Jun-Young Cha, Hyung-In Yoon, In-Sung Yeo, Kyung-Hoe Huh, and Jung-Suk Han. Panoptic segmentation on panoramic radiographs: Deep learning-based segmentation of various structures including maxillary sinus and mandibular canal. *Journal of Clinical Medicine*, 10(12):2577, 2021. 1

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 2

[6] Liang-Chieh Chen, Raphael Gontijo Lopes, Bowen Cheng, Maxwell D Collins, Ekin D Cubuk, Barret Zoph, Hartwig Adam, and Jonathon Shlens. Naive-student: Leveraging semi-supervised learning in video sequences for urban scene segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 695–714. Springer, 2020. 1, 2

[7] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12475–12485, 2020. 1, 2

[8] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34:17864–17875, 2021. 1, 2, 4, 7

[9] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 1, 2, 3, 4, 5, 7

[10] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023. 2, 3, 4, 5

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1

[12] Osmar Luiz Ferreira de Carvalho, Osmar Abílio de Carvalho Júnior, Cristiano Rosa e Silva, Anesmar Olino de Albuquerque, Nickolas Castro Santana, Dibio Leandro Borges, Roberto Arnaldo Trancoso Gomes, and Renato Fontes Guimarães. Panoptic segmentation meets remote sensing. *Remote Sensing*, 14(4):965, 2022. 1

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[14] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11583–11592, 2022. 2

[15] 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. 2021. 4

[16] Geoffrey French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham D. Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*, 2020. 2, 7

[17] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European Conference on Computer Vision*, pages 540–557. Springer, 2022. 2, 4, 5

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[19] Ivan Grubišić, Marin Oršić, and Siniša Šegvić. Revisiting consistency for semi-supervised semantic segmentation. *Sensors*, 23(2):940, 2023. 2

[20] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2022. 2, 5

[21] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2

[22] Lukas Hoyer, David Joseph Tan, Muhammad Ferjad Naeem, Luc Van Gool, and Federico Tombari. Semivl: Semi-supervised semantic segmentation with vision-language guidance. In *European Conference on Computer Vision*, pages 257–275. Springer, 2025. 1, 2, 4, 5, 7, 3, 6

[23] Hanzhe Hu, Fangyun Wei, Han Hu, Qiwei Ye, Jinshi Cui, and Liwei Wang. Semi-supervised semantic segmentation via adaptive equalization learning. *Advances in Neural Information Processing Systems*, 34:22106–22118, 2021. 2, 7

[24] Jie Hu, Chen Chen, Liujuan Cao, Shengchuan Zhang, Annan Shu, Guannan Jiang, and Rongrong Ji. Pseudo-label alignment for semi-supervised instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16337–16347, 2023. 1

[25] Xinting Hu, Li Jiang, and Bernt Schiele. Training vision transformers for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4007–4017, 2024. 2, 7

[26] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2

[27] Siyu Jiao, Yunchao Wei, Yaowei Wang, Yao Zhao, and Humphrey Shi. Learning mask-aware clip representations for zero-shot segmentation. *Advances in Neural Information Processing Systems*, 36:35631–35653, 2023. 2

[28] Rihuan Ke, Angelica I Aviles-Rivero, Saurabh Pandey, Saikumar Reddy, and Carola-Bibiane Schönlieb. A three-stage self-training framework for semi-supervised semantic segmentation. *IEEE Transactions on Image Processing*, 31: 1805–1815, 2022. 2

[29] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 5

[30] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6399–6408, 2019. 1

[31] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9404–9413, 2019. 1, 2, 5

[32] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2, 5, 7, 3

[33] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. Open-vocabulary object detection upon frozen vision and language models. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[34] Wei-Sheng Lai, Jia-Bin Huang, and Ming-Hsuan Yang. Semi-supervised learning for optical flow with generative adversarial networks. *Advances in neural information processing systems*, 30, 2017. 1

[35] Xin Lai, Zhuotao Tian, Li Jiang, Shu Liu, Hengshuang Zhao, Liwei Wang, and Jiaya Jia. Semi-supervised semantic segmentation with directional context-aware consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1205–1214, 2021. 2

[36] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. In *ECCV*, 2024. 2, 4

[37] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022. 2

[38] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3041–3050, 2023. 3

[39] Qizhu Li, Anurag Arnab, and Philip HS Torr. Weakly-and semi-supervised panoptic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 102–118, 2018. 1

[40] Chen Liang, Wenguan Wang, Jiaxu Miao, and Yi Yang. Logic-induced diagnostic reasoning for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16197–16208, 2023. 2, 7

[41] 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. 2

[42] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1, 5

[43] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on com-*

puter vision and pattern recognition, pages 11976–11986, 2022. 2, 5

[44] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 5

[45] Huayu Mai, Rui Sun, Tianzhu Zhang, and Feng Wu. Rankmatch: Exploring the better consistency regularization for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3391–3401, 2024. 2

[46] Ivan Martinović, Josip Šarić, and Siniša Šegvić. Mc-panda: Mask confidence for panoptic domain adaptation. In *European Conference on Computer Vision*, pages 167–185. Springer, 2024. 2

[47] Sudhanshu Mittal, Maxim Tatarchenko, and Thomas Brox. Semi-supervised semantic segmentation with high-and low-level consistency. *IEEE transactions on pattern analysis and machine intelligence*, 43(4):1369–1379, 2019. 2

[48] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12674–12684, 2020. 1, 2

[49] Lu Qi, Jason Kuen, Zhe Lin, Jiuxiang Gu, Fengyun Rao, Dian Li, Weidong Guo, Zhen Wen, Ming-Hsuan Yang, and Jiaya Jia. Ca-ssl: Class-agnostic semi-supervised learning for detection and segmentation. In *European Conference on Computer Vision*, pages 59–77. Springer, 2022. 1, 2

[50] Mengshi Qi, Yunhong Wang, Jie Qin, and Annan Li. Kegan: Knowledge embedded generative adversarial networks for semi-supervised scene parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5237–5246, 2019. 2

[51] Siyuan Qiao, Wei Shen, Zhishuai Zhang, Bo Wang, and Alan Yuille. Deep co-training for semi-supervised image recognition. In *Proceedings of the european conference on computer vision (eccv)*, pages 135–152, 2018. 2

[52] 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. 1, 2, 3, 4, 5, 7

[53] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 2

[54] Tal Ridnik, Emanuel Ben Baruch, Asaf Noy, and Lihi Zelnik. Imagenet-21k pretraining for the masses. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021. 2

[55] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[56] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10765–10775, 2021. 1

[57] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 4, 5, 2

[58] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 1, 2, 3

[59] Nasim Souly, Concetto Spampinato, and Mubarak Shah. Semi supervised semantic segmentation using generative adversarial network. In *Proceedings of the IEEE international conference on computer vision*, pages 5688–5696, 2017. 2

[60] Boyuan Sun, Yuqi Yang, Le Zhang, Ming-Ming Cheng, and Qibin Hou. Corrmatch: Label propagation via correlation matching for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3097–3107, 2024. 2

[61] 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. 2, 3, 5

[62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. 2

[63] Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. In *European Conference on Computer Vision*, pages 315–332. Springer, 2024. 2, 4

[64] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5463–5474, 2021. 2, 3

[65] Haonan Wang, Qixiang Zhang, Yi Li, and Xiaomeng Li. Allspark: Reborn labeled features from unlabeled in transformer for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3627–3636, 2024. 2, 7

[66] Xiaoyang Wang, Huihui Bai, Limin Yu, Yao Zhao, and Jimin Xiao. Towards the uncharted: Density-descending feature perturbation for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3303–3312, 2024. 2

[67] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gon-

tijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7959–7971, 2022. 2

[68] Linshan Wu, Leyuan Fang, Xingxin He, Min He, Jiayi Ma, and Zhun Zhong. Querying labeled for unlabeled: Cross-image semantic consistency guided semi-supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8827–8844, 2023. 2, 7

[69] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268, 2020. 2, 3

[70] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 1, 2, 4, 5, 7

[71] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3060–3069, 2021. 1

[72] Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *CVPR*, 2023. 2

[73] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. St++: Make self-training work better for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4268–4277, 2022. 2

[74] Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang, and Yinghuan Shi. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7236–7246, 2023. 2, 5, 6, 7

[75] Qihang Yu, Huiyu Wang, Siyuan Qiao, Maxwell Collins, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. k-means mask transformer. In *European Conference on Computer Vision*, pages 288–307. Springer, 2022. 2, 3

[76] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *Advances in Neural Information Processing Systems*, 36:32215–32234, 2023. 1, 2, 4, 5

[77] Jianlong Yuan, Yifan Liu, Chunhua Shen, Zhibin Wang, and Hao Li. A simple baseline for semi-supervised semantic segmentation with strong data augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8229–8238, 2021. 2

[78] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 5, 3

[79] Oliver Zendel, Matthias Schörghuber, Bernhard Rainer, Markus Murschitz, and Csaba Beleznai. Unifying panoptic segmentation for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21351–21360, 2022. 1

[80] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 2

[81] Yuanyi Zhong, Bodi Yuan, Hong Wu, Zhiqiang Yuan, Jian Peng, and Yu-Xiong Wang. Pixel contrastive-consistent semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7273–7282, 2021. 2, 7

[82] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 5

[83] 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. 2, 4

[84] Yuliang Zou, Zizhao Zhang, Han Zhang, Chun-Liang Li, Xiao Bian, Jia-Bin Huang, and Tomas Pfister. Pseudoseg: Designing pseudo labels for semantic segmentation. In *International Conference on Learning Representations*, 2021. 2, 7

[85] Lojze Žust, Janez Perš, and Matej Kristan. Lars: A diverse panoptic maritime obstacle detection dataset and benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20304–20314, 2023. 1