Robust Pseudo-Labeling via Decoupled Class-Aware Filtering and Dynamic Category Correction

Jianghang Lin[1], Yilin Lu[1], Chaoyang Zhu[1], Yunhang Shen[1], Shengchuan Zhang[1*], Liujuan Cao[1]

[1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University, 361005, P.R. China.
{hunterjlin007, yilinlu}@stu.xmu.edu.cn, {seanzhuh, shenyunhang01}@gmail.com, {zsc_2016, caoliujuan}@xmu.edu.cn

Abstract

Semi-Supervised Instance Segmentation (SSIS) involves classifying and grouping image pixels into distinct object instances using limited labeled data alongside large-scale unlabeled data. A major challenge in SSIS lies in the inherent noise of pseudo-labels, particularly when class and mask qualities are coupled into a single confidence score for filtering. Such coupling often results in sub-optimal trade-offs between semantic accuracy and spatial precision. To address this, we propose a novel **P**seudo-**L**abel **D**ecoupling and **C**orrection (**PL-DC**) framework, which explicitly decouples and enhances the pseudo-label selection process for SSIS. At the instance level, we introduce a Decoupled Filtering with Adaptive Class-Aware Thresholds mechanism, which independently evaluates class and mask qualities using category-specific thresholds updated via exponential moving averages. At the category level, we design a Dynamic Instance Category Correction module that reassigns ambiguous class pseudo-label by leveraging semantic prototypes and consistency alignment. At the pixel level, a Pixel-Level Mask Uncertainty-Aware mechanism is applied to suppress the influence of unreliable pixels during mask supervision, further improving the robustness against pixel-wise noise. Extensive experiments on COCO and Cityscapes datasets demonstrate that the proposed PL-DC achieves significant performance improvements, setting new state-of-the-art results. Notably, PL-DC achieves gains of $+11.7$ mAP with just 1% labeled COCO data and $+16.4$ mAP with 5% Cityscapes labels, showing its effectiveness under extremely low-label regimes.

Code — https://github.com/HunterJ-Lin/PL-DC
Extended version — https://arxiv.org/abs/2505.11075

Introduction

Instance segmentation has achieved remarkable progress in recent years, largely driven by the success of deep learning and the availability of large-scale annotated datasets such as COCO (Lin et al. 2014), LVIS (Gupta, Dollár, and Girshick 2019), Cityscapes (Cordts et al. 2016), and BDD100K (Yu et al. 2020). These datasets have enabled extensive research in fully-supervised instance segmentation (FSIS), where each pixel is densely labeled and associated with object instance IDs. Nevertheless, the laborious and lavish collection of pixel-level annotations has severely barricaded the applicability of FSIS in practical application. Semi-supervised instance segmentation (SSIS) has thus emerged as a promising direction that seeks to reduce annotation dependency by leveraging a small labeled set along with abundant unlabeled data. While semi-supervised learning has made significant strides in classification and detection tasks, SSIS remains particularly challenging due to its joint requirement for semantic classification, instance-level separation, and pixel-level mask prediction. Through systematic analysis, we identify three core challenges that hinder the advancement of SSIS: **(1) Instance-level filtering bias.** Most existing methods (Filipiak et al. 2024; Berrada et al. 2024) filter pseudo-labels based on a single confidence score that couples class confidence and mask quality. However, as shown in Fig. 2 (a), this coupled score is poorly correlated with true instance quality, leading to bias estimation of pseudo-labels. **(2) Category-level confusion.** As shown in Fig. 3, classes with high visual similarity or frequent co-

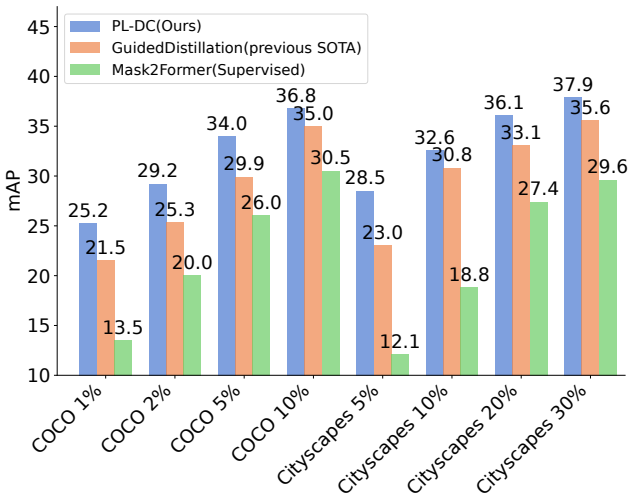

Figure 1: The proposed PL-DC outperforms the previous state-of-the-art SSIS method, GuidedDistillation (Berrada et al. 2024), across all settings. Moreover, PL-DC achieves significant improvements compared to the fully-supervised Mask2Former (Cheng et al. 2022).

*Corresponding Author

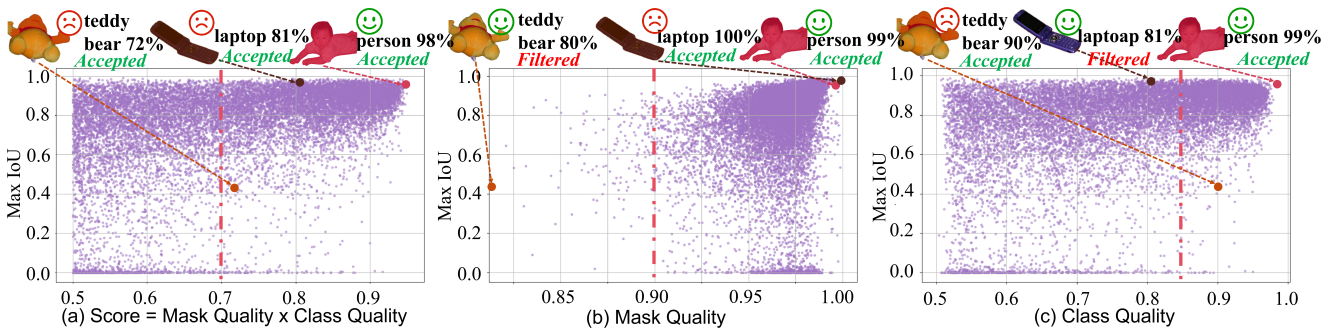

Figure 2: Relationship between predicted instance scores and IoU. (a) Combined instance scores show weak correlation with IoU. (b) Mask quality and (c) Class quality independently influence pseudo-label segmentation and classification.

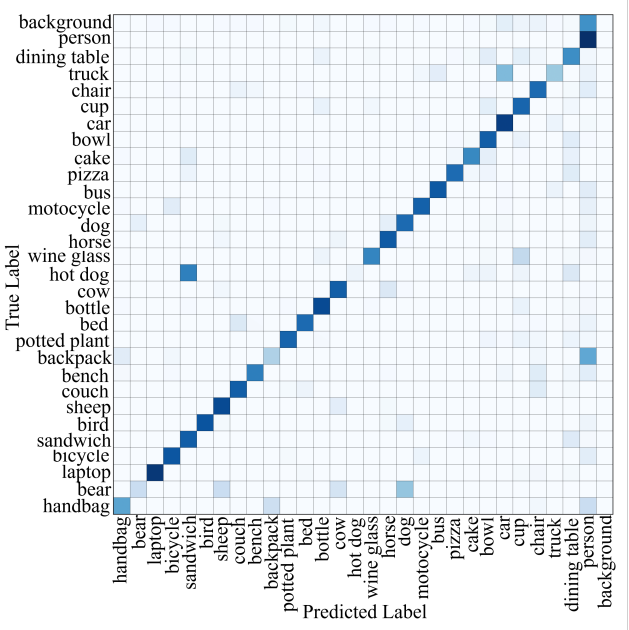

Figure 3: Confusion matrix of the model trained on 1% COCO. For clarity, we visualize only the 29 most confused object categories and 1 background category.

occurrence (e.g., *Bears* vs. *Dogs*, *Hot Dogs* vs. *Sandwiches*) often confuse the model, resulting in ambiguous or incorrect category predictions. This is exacerbated in the semi-supervised setting where category supervision is sparse. **(3) Pixel-level noise accumulation.** Compared to instance category predictions, mask pseudo-labels are much denser and noisier. Because pixel-level mask losses are computed over large regions, noisy masks can easily dominate the training signal and harm model optimization.

To address the above challenges, we propose a new SSIS framework called **P**seudo-**L**abel **D**ecoupling and **C**orrection (**PL-DC**), which integrates three novel components: **(1) Decoupled Filtering with Adaptive Class-Aware Thresholds.** Motivated by the observation from Fig. 2 (b)(c) that class and mask quality influence different aspects of instance quality, we decouple the filtering process by assigning separate thresholds to class and mask confidence. These thresholds are updated dynamically via exponential moving averages based on the per-category distribution of pseudo-label scores. This filtering strategy avoids the over-reliance on a single coupled score and enables fine-grained, class-aware control over pseudo-label selection, thus eliminating the detrimental effects of potential trade-offs between class and mask quality. **(2) Dynamic Instance Category Correction.** To mitigate category-level confusion, we introduce a correction module that leverages the pretrained visual-language model CLIP to refine the category distribution of filtered pseudo-labels. By computing similarity between masked instance region features and enhanced textual category descriptions, we recalibrate the predicted class probabilities to align better with semantic consistency. **(3) Pixel-Level Mask Uncertainty-Aware.** We apply a pixel-level loss weighting strategy that down-weights uncertain regions and emphasizes confident pixels, reducing noisy pseudo-label influence and improving mask learning robustness.

Extensive experiments on COCO and Cityscapes demonstrate that our PL-DC achieves new state-of-the-art results. Specifically, on COCO with 1%, 2%, 5%, 10%, and 100% labeled images, our proposed PL-DC achieves significant performance boosts with increases of +11.7%, +9.2%, +8.0%, +6.3%, and +5.3% mAP, respectively. On the Cityscapes dataset, with 5%, 10%, 20%, and 30% labeled images, PL-DC also exhibits substantial enhancements, recording mAP improvements of +16.4%, +13.8%, +8.7%, and +8.3%, respectively.

Related Work

Instance Segmentation

Instance segmentation assigns a category label to each pixel while distinguishing individual object instances. Existing approaches fall into three main families: detection-based, clustering-based, and query-based. *Detection-based methods* (He et al. 2017; Bolya et al. 2019; Li et al. 2022) extend object detectors by predicting masks within detected regions. Mask R-CNN (He et al. 2017) remains a representative framework, building on Faster R-CNN (Girshick 2015) and widely used as a strong baseline. *Clustering-*

based methods (Gao et al. 2019; Liu et al. 2017; Bai and Urtasun 2017) group pixels based on feature similarity and spatial affinity. Techniques like Mean Shift (Comaniciu and Meer 2002) or Graph Cut (Boykov and Jolly 2001) are commonly used for pixel clustering. A representative method is the Deep Watershed Transform (Bai and Urtasun 2017) which views an image as a topographic surface and apply watershed-based grouping to form instances. *Query-based methods* leverage learnable queries, typically with transformer architectures, to directly predict instance masks. DETR (Carion et al. 2020) introduces set-based bipartite matching for end-to-end instance prediction, while MaskFormer (Cheng, Schwing, and Kirillov 2021) and Mask2Former (Cheng et al. 2022) unify semantic, instance, and panoptic segmentation within one framework.

Semi-Supervised Instance Segmentation

Semi-supervised learning reduces reliance on labeled data by incorporating unlabeled images, and has been widely explored in image classification and object detection via pseudo-labeling, consistency regularization, and adversarial training. Pseudo-label-based approaches (Lin et al. 2024; Xu et al. 2021; Mi et al. 2022) use a teacher model to annotate unlabeled data, while consistency-based methods (Berthelot et al. 2019a,b; Gao et al. 2020; Jeong et al. 2019a) enforce prediction stability under perturbations. For instance segmentation, the pixel-level nature of masks makes noisy pseudo-labels particularly challenging. Noisy Boundary (Wang, Li, and Wang 2022) was the first to formally introduce the semi-supervised instance segmentation task. Noisy Boundary (Wang, Li, and Wang 2022) first formalized semi-supervised instance segmentation, modeling label noise near object boundaries and using noise-tolerant mask prediction and boundary-aware maps. Polite Teacher (Filipiak et al. 2024) adopts dynamic pseudo-labeling within a teacher–student framework using CenterMask (Lee and Park 2020), relying on confidence thresholding to filter noisy pseudo-labels. PAIS (Hu et al. 2023) instead reweights semi-supervised losses using dynamic class–mask score pairs rather than discarding low-confidence pseudo-labels. More recently, GuidedDistillation (Berrada et al. 2024) applies a three-stage teacher–student distillation pipeline built on the query-based Mask2Former (Cheng et al. 2022), achieving strong performance but still limited by coupled score filtering of pseudo-labels.

Method

In this section, we outline our PL-DC framework designed to tackle the three challenges commonly encountered in semi-supervised instance segmentation, as depicted in Fig. 4. The objective is to leverage both labeled data $D_L = \{X_L, Y_L\}$ and unlabeled data $D_U = \{X_U\}$ to optimize instance segmentation performance, where X represents image samples and Y denotes mask annotations with their corresponding classes. Our framework utilizes a teacher-student structure in semi-supervised learning. It incorporates two instance segmentation networks with identical structures: one acting as the teacher and the other as the student. The teacher generates pseudo-labels for the unlabeled data, which the student uses to learn alongside the labeled data. Consequently, the overarching loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \lambda \mathcal{L}_{\text{unsup}}, \tag{1}$$

where $\mathcal{L}_{\text{sup}}$ and $\mathcal{L}_{\text{unsup}}$ denote the losses for supervised and unsupervised learning, respectively, and λ is a hyper-parameter that balances these losses. For instance segmentation, the supervised learning loss is defined as:

$$\mathcal{L}_{\text{sup}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{mask}}, \tag{2}$$

with $\mathcal{L}_{\text{cls}}$ is the classification cross-entropy loss and $\mathcal{L}_{\text{mask}}$ is the pixel-level binary cross-entropy loss, which may optionally include dice loss. The unsupervised learning loss mirrors Eq. 2, but the supervision comes from the pseudo-labels generated by the teacher. The student updates its parameters via stochastic gradient descent (SGD). To prevent overfitting, the teacher's gradients are frozen, and its parameters are updated from the student using the Exponential Moving Average (EMA) (Tarvainen and Valpola 2017).

Decoupled Filtering with Adaptive Class-Aware Thresholds

We take Mask2Former (Cheng et al. 2022) as our foundational instance segmentation network structure due to its strong performance in this field. This model computes the instance score $\mathbf{s}_k$ as the product of the class quality $\mathbf{c}_k$ and mask quality $\mathbf{m}_k$. Class quality $\mathbf{c}_k$ is defined as:

$$\mathbf{c}_k = \frac{e^{x_k^j}}{\sum_{i=1}^{N} e^{x_k^i}}, \tag{3}$$

where N is the number of class, x_k^i represents the logit of class i prediction for the k-th instance, and x_k^j is the maximum class logit. Mask quality $\mathbf{m}_k$ is calculated as:

$$\mathbf{m}_k = \frac{\sum_{i=1}^{HW} \sigma(q_k^i)\mathbf{1}[\sigma(q_k^i) > 0.5]}{\sum_{i=1}^{HW} \mathbf{1}[\sigma(q_k^i) > 0.5]}, \tag{4}$$

where HW is the total number of pixels in the mask, q_k^i is the per-pixel logit of the predicted mask for the k-th instance, σ denotes the sigmoid function, and $\mathbf{1}[\sigma(q_k^i) > 0.5]$ is an indicator function that equals 1 if $\sigma(q_k^i) > 0.5$, and 0 otherwise. In fully-supervised learning settings, where abundant labeled data are available, the model can effectively optimize both $\mathbf{c}_k$ and $\mathbf{m}_k$, allowing the combined instance score $\mathbf{s}_k = \mathbf{c}_k \cdot \mathbf{m}_k$ to serve as a reliable metric of pseudo-label quality. However, in semi-supervised learning, the limited quantity of labeled data and the inevitable presence of noisy pseudo-labels from unlabeled data hinder the simultaneous optimization of $\mathbf{c}_k$ and $\mathbf{m}_k$. This constraint often results in the coupled instance score failing to accurately reflect the actual reliability of a prediction. Using a single static threshold on $\mathbf{s}_k$ to filter pseudo-labels, as in prior works (Berrada et al. 2024), imposes a competitive relationship between class and mask quality. For example, a score of $\mathbf{s}_k$ of 0.72 could emerge from either a high mask quality $\mathbf{m}_k$ of 0.96 combined with a lower class quality $\mathbf{c}_k$ of 0.75, or from more balanced values of 0.8 and 0.9, respectively.

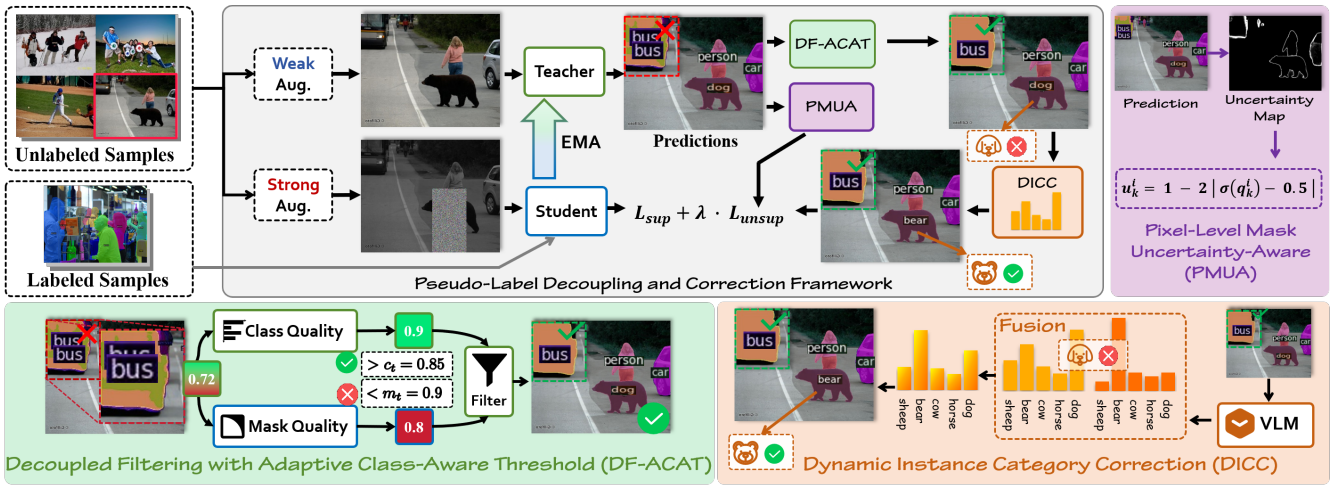

Figure 4: Framework of our proposed pseudo-label decoupling and correction (PL-DC) for semi-supervised instance segmentation. PL-DC includes two segmentation models, both Mask2Former (Cheng et al. 2022), with identical configurations, namely Teacher and Student. The Teacher model generates an uncertainty map for Pixel-Level Mask Uncertainty-Aware training, filters pseudo-labels by the Decoupled Filtering with Adaptive Class-Aware Thresholds (DF-ACAT) mechanism, and further corrects category by Dynamic Instance Category Correction (DICC). The Teacher's parameters are gradually updated from the Student model via Exponential Moving Average (EMA). The Student is trained using both ground-truth labels and pseudo-labels (with uncertainty map), denoted as $\mathcal{L}_{\text{sup}}$ and $\mathcal{L}_{\text{unsup}}$, respectively.

In such cases, filtering decisions based only on the overall score can lead to the retention of pseudo-labels with either insufficient semantic correctness or poor spatial delineation.

We observe that the decoupled class quality $\mathbf{c}_k$ and mask quality $\mathbf{m}_k$ independently control the quality of instance pseudo-labels, as illustrated in Fig. 2 (b)(c). Based on this insight, we propose a Decoupled Filtering with Adaptive Class-Aware Thresholds (DF-ACAT) mechanism that jointly incorporates class-aware adaptation, progressive threshold scheduling, and fine-grained multi-level filtering to better align pseudo-label selection with evolving model confidence. Specifically, we decouple the class quality c_k and mask quality m_k as independent criteria for filtering, reducing the bias introduced by coupling. Instead of applying fixed thresholds, we initialize per-class thresholds for both class and mask quality based on the frequency of labeled samples. Specifically, given a labeled set with class counts $\{n_1, n_2, ..., n_K\}$ and total count $N = \sum_k n_k$, the initial threshold for class k is defined as:

$$\mathcal{T}_c^{(k,0)} = \mathcal{T}_c^{min} + (\mathcal{T}_c^{max} - \mathcal{T}_c^{min}) \cdot \frac{n_k}{N}, \qquad (5)$$

and similarly for $\mathcal{T}_m^{(k,0)}$, where $\mathcal{T}_c^{min}$ and $\mathcal{T}_c^{max}$ are hyper-parameters. This ensures that common classes start with stricter thresholds, while rare classes are treated more permissively. During training, the teacher model processes weakly augmented unlabeled image X_U^{weak} and produces Q candidate instance predictions $\{(c_i, m_i)\}_{i=1}^{Q}$. Unlike static thresholding, we adopt a dynamic update strategy using an EMA over per-class pseudo-label scores to adjust thresholds at each epoch. Let $\widehat{\mathcal{T}}_c^{(k)}(e)$ denote the average class quality score of predicted pseudo-labels for class k at epoch e, then the updated threshold is:

$$\mathcal{T}_c^{(k)}(e) = \alpha \cdot \mathcal{T}_c^{(k)}(e-1) + (1-\alpha) \cdot \widehat{\mathcal{T}}_c^{(k)}(e), \qquad (6)$$

and similarly for $\mathcal{T}_m^{(k)}(e)$, where α is the EMA smoothing coefficient the same as teacher model parameters update. This formulation allows the thresholds to be updated stably and responsively, without requiring handcrafted schedules. We then categorize pseudo-labels into three levels based on these thresholds: those exceeding both $\mathcal{T}_c^{(k)}$ and $\mathcal{T}_m^{(k)}$ are considered high-quality and directly retained for training; those far below are discarded; and those in between are passed to the **Dynamic Instance Category Correction (DICC)** module for further refinement. This stratified filtering strategy ensures that high-quality supervision signals are exploited early, while ambiguous cases are progressively resolved with the assistance of additional guidance.

Dynamic Instance Category Correction

Ideally, semi-supervised learning alleviates label scarcity, but it is often hindered by severe class imbalance in instance segmentation. For instance, in the COCO dataset, *person* instances make up 30% of all foreground training instances, while *hair driers* and *toasters* represent only 0.023% and 0.026%, respectively. This imbalance biases the model toward predicting dominant classes, especially when training data is limited, further exacerbating skewed pseudo-label generation. As depicted in Fig. 3, instances with similar appearances or frequent co-occurrences are particularly prone to misclassification into dominant categories. For example, bears are often confused with dogs due to visual similarity, and hot dogs are frequently mislabeled as sandwiches due to contextual co-occurrence.

Recently, large visual-language alignment models (VLM) pre-trained on extensive image-text pairs have demonstrated strong zero-shot classification capabilities. These models can effectively address the inaccuracies in pseudo-labels. To leverage this advantage, we propose Dynamic Instance Category Correction (DICC) to rectify the categories of medium-quality pseudo-labels after DF-ACAT. For simplicity, we utilize CLIP (Radford et al. 2021) as a representative VLM in our DICC. Given each medium-quality pseudo-label (C_i, M_i), the most direct way is to extract the image patch x_i^{pool} cropped by the mask M_i. CLIP then computes the similarity between this patch and the textual descriptions of all categories $\mathbf{t} \in R^N$, producing a probability distribution $p_i^{clip} \in R^N$:

$$p_i^{clip} = \text{Softmax}(\text{CLIP}_V(x_i^{pool}) \cdot \text{CLIP}_T(\mathbf{t})), \quad (7)$$

where CLIP_V and CLIP_T denote CLIP's vision and text encoders, respectively. However, this approach is computationally intensive and loses background information, which is critical for global alignment. To address this, we introduce a more efficient batching strategy. For a given weakly unlabeled image X_U^{weak}, we extract a global feature V_g using CLIP's visual backbone. Pooling V_g over all masks yields mask embeddings $E_m \in R^{M,d}$. We then compute the similarity between E_m and the text embeddings of all categories to obtain the category probabilities. To improve category discrimination, we further augment each category description using a large language model (e.g., GPT-4o) to generate k diverse descriptions per category, constructing enhanced textual representations t_e. Equ. 7 is thus reformulated as:

$$t_e^i = \sum_{j=1}^{k} \text{Softmax}(\text{CLIP}_T(t^i_{e_1,\ldots,e_k})) \cdot t^i_{e_1,\ldots,e_k}, \quad (8)$$

$$p_i^{clip} = \text{Softmax}(E_m \cdot \{t_e^1, \ldots, t_e^N\}). \quad (9)$$

Next, we dynamically fuse the teacher model's prediction $\hat{p}_i$ with p_i^{clip} to obtain a final probability distribution $p_i^f \in R^N$. The category with the highest score in p_i^f is selected as the corrected category pseudo-label C_i^{corr}:

$$w = 0.25(\cos(\frac{e}{E_{max}}\pi) + 1), \quad (10)$$

$$p_i^f = w \cdot p_i^{clip} + (1 - w) \cdot \hat{p}_i, \quad (11)$$

$$C_i^{corr} = \arg\max(p_i^f), \quad (12)$$

where e and E_{max} denote the current and maximum training epochs, respectively. The weighting factor w decays from 0.5 to 0 over training, reflecting the increasing reliability of the teacher model as learning progresses. This dynamic strategy effectively balances the strengths of both the teacher model and the VLM, enabling complementary recognition of challenging or unfamiliar categories.

Pixel-Level Mask Uncertainty-Aware

In instance segmentation, the loss function for model training typically includes an instance-level classification cross-entropy loss and a pixel-level mask binary cross-entropy loss. The pixel-level mask loss considers all pixels in the entire image, whereas the instance-level classification loss is concerned with a relatively smaller number of instances. Consequently, the number of pixel-level mask pseudo-labels significantly exceeds that of instance-level category pseudo-labels, making the model training more susceptible to the influence of pixel-level mask pseudo-labels. Given the extensive use of pixel-level mask pseudo-labels in semi-supervised learning, it is crucial to account for the uncertainty associated with these labels.

Recent work Noisy Boundaries (Wang, Li, and Wang 2022) introduced the Boundary-preserving Map (BPM), which re-weights the mask loss for different pixels based on their proximity to object boundaries, thereby making model training sensitive to uncertain mask pixels. Noisy Boundaries posits that uncertainty primarily exists at object boundaries. However, we have observed significant uncertainty in areas where multiple objects overlap, a scenario where BPM is less effective. To address this broader range of uncertainties, we propose the simple yet effictive Pixel-level Mask Uncertainty-Aware (PMUA) approach to re-weight the mask loss across different pixels comprehensively. We define the uncertainty u_k^i of the per-pixel mask as:

$$u_k^i = 1 - 2\left|\sigma(q_k^i) - 0.5\right|, \quad (13)$$

where $\sigma(q_k^i)$ is the predicted foreground per-pixel binary mask probability of the k-th instance by the teacher model. Following the DF-ACAT and DICC processes, we obtain corrected pseudo-labels $\hat{Y}_U^{corr} = \{(C_k^{corr}, M_k, u_k) \mid C_k^{corr} \in \{1, \ldots, N\}, M_k \in \{0,1\}^{HW}, u_k \in [0,1]^{HW}\}_{k=0}^{N^{pgt}}$ for unlabeled data, where C_k^{corr} is the corrected pseudo ground truth class labels, M_k is the pseudo ground truth binary mask, and u_k represents the uncertainty values for each M_k, N^{pgt} is the total number of pseudo-labeled instances obtained. Then, pixel-level mask binary cross-entropy loss for unlabeled data to train the student model is defined as:

$$\begin{aligned}\mathcal{L}_{mask}^{unsup} = &-\frac{1}{QHW}\sum_{k=1}^{Q}\sum_{i=1}^{HW}(1 - u^i_{\hat{\sigma}(k)})[M^i_{\hat{\sigma}(k)}\log(t_k^i) \\ &+(1 - M^i_{\hat{\sigma}(k)})\log(1 - t_k^i)],\end{aligned} \quad (14)$$

where $\hat{\sigma}$ is the optimal assignment calculated using the Hungarian algorithm, t_k^i is the predicted foreground per-pixel binary mask probability of the k-th instance by student model. In the **Extended version**, we derive the gradient of $\mathcal{L}_{mask}^{unsup}$ with respect to the student model parameters θ and prove that a higher $u^i_{\hat{\sigma}(k)}$, indicating greater noise in $M^i_{\hat{\sigma}(k)}$, results in a proportionally lesser influence of the pixel's pseudo-label on the update of θ, thereby improving the robustness of the training process under label uncertainty.

Experiments

Settings and Implementation Details

Experimental Settings. We benchmark our proposed PLDC on COCO (Lin et al. 2014) and Cityscapes (Cordts et al.

Method	1%	2%	5%	10%	100%
Mask-RCNN, Superised	3.5	9.3	17.3	22.0	34.5
Mask2Former, Superised	13.5	20.0	26.0	30.5	43.5
DD (Radosavovic et al. 2018)	3.8	11.8	20.4	24.2	35.7
Noisy Boundaries (Wang, Li, and Wang 2022)	7.7	16.3	24.9	29.2	38.6
Polite Teacher (Filipiak et al. 2024)	18.3	22.3	26.5	30.8	-
PAIS (Hu et al. 2023)	21.1	-	29.3	31.0	39.5
GuidedDistillation (Berrada et al. 2024)	21.5	25.3	29.9	35.0	-
PL-DC (Ours)	**25.2**	**29.2**	**34.0**	**36.8**	**48.8**

Table 1: Comparison with other SSIS on COCO.

Method	5%	10%	20%	30%
Mask-RCNN, Supervised	11.3	16.4	22.6	26.6
Mask2Former, Supervised	12.1	18.8	27.4	29.6
DD (Radosavovic et al. 2018)	13.7	19.2	24.6	27.4
STAC (Sohn et al. 2020)	11.9	18.2	22.9	29.0
CSD (Jeong et al. 2019b)	14.1	17.9	24.6	27.5
CCT (Ouali, Hudelot, and Tami 2020)	15.2	18.6	24.7	26.5
Dual-branch (Luo and Yang 2020)	13.9	18.9	24.0	28.9
Ubteacher (Liu et al. 2021)	16.0	20.0	27.1	28.0
Noisy Boundaries (Wang, Li, and Wang 2022)	17.1	22.1	29.0	32.4
PAIS (Hu et al. 2023)	18.0	22.9	29.2	32.8
GuidedDistillation (Berrada et al. 2024)	23.0	30.8	33.1	35.6
PL-DC (Ours)	**28.5**	**32.6**	**36.1**	**37.9**

Table 2: Comparison with other SSIS on Cityscapes.

2016) datasets following existing works (Berrada et al. 2024; Hu et al. 2023; Wang, Li, and Wang 2022). The COCO dataset, which comprises 80 categories, is notably challenging for instance segmentation. It includes $118k$ *train2017* labeled images, $5k$ *val2017* labeled images and $123k$ *unlabel2017* unlabeled images. We randomly sample 1%, 2%, 5%, and 10% of the images from the *train2017* split as labeled data and treated the rest as unlabeled data following common settings. Additionally, we utilized the entire *train2017*, denoted as 100%, as labeled data and incorporated the *unlabel2017* as unlabeled data for PL-DC evaluation. The Cityscapes dataset contains $2,975$ training images and 500 validation images of size 1024×2048 taken from a car driving in German cities, labeled with 8 semantic instance categories. We follow (Berrada et al. 2024) sample 5%, 10%, 20%, and 30% of the images from the training set as labeled images and treat the rest as unlabeled ones. We conducted evaluations using the COCO *val2017* and the Cityscapes validation sets for their respective experimental settings, reporting the standard COCO mAP metric.

Implementation Details. We employ Mask2Former (Cheng et al. 2022) with ResNet-50 as our baseline instance segmentation network, and the implementation and hyper-parameters setting are the same as those in Detectron2 (Wu et al. 2019). By default, all experiments are conducted on a single machine equipped with eight 4090 GPUs, each with 24 GB of memory. For optimization, we utilize AdamW (Loshchilov and Hutter 2017) with a learning rate and weight decay both set at 0.0001. Due to limited GPU memory, all backbones are frozen. Following (Liu et al. 2021), we apply random horizontal flip and scale jittering as weak augmentations for the teacher model, while the student model receives strong augmentations including horizontal flip, scale jittering, color jittering, grayscale, gaussian blur, and CutOut (DeVries and Taylor 2017). We use $\mathcal{T}_c^{min} = 0.7$, $\mathcal{T}_c^{max} = 0.85$, $\mathcal{T}_m^{min} = 0.8$ and $\mathcal{T}_m^{max} = 0.9$ for DF-ACAT. We use $\alpha = 0.9996$ for EMA and $\lambda = 1$ for the unsupervised loss $\mathcal{L}_{\text{unsup}}$. For the COCO setup, we pretrain the teacher model with the supervised learning defined in Equ. 2 about 20k iterations. Afterward, the student model is initialized with the parameters of the teacher model. The total training iterations for each semi-supervised learning are all $360K$ (50 epochs), with batch sizes consistently comprising 8 labeled and 8 unlabeled images unless otherwise specified. For Cityscapes setup, the hyper-parameters mirror those of the COCO configuration, except the total training duration is reduced to $180k$ iterations.

Comparison with Other Methods

In Tab. 1, We compare our PL-DC with other semi-supervised instance segmentation frameworks on the COCO dataset. Our observations reveal that PL-DC consistently outperforms the current state-of-the-art method, GuidedDistillation (Berrada et al. 2024), across all COCO-labeled data ratios. Notably, our PL-DC shows a more substantial increase in mAP at lower labeled data ratios compared to the fully supervised Mask2Former. Specifically, the mAP improvements are $+11.7$, $+9.2$, $+8.0$, and $+6.3$ for 1%, 2%, 5%, and 10% labeled data, respectively, underscoring PL-DC's effective use of large-scale unlabeled data. In contrast, GuidedDistillation exhibits smaller and somewhat counterintuitive mAP gains of $+3.9$ at 5% and $+4.5$ at 10%, indicating a higher dependency on labeled data. Moreover, employing 100% of the COCO labeled data, PL-DC further achieves an enhancement of $+5.3$ mAP by integrating $123k$ *unlabel2017* COCO images.

To evaluate the generalizability of our PL-DC, we conducted experiments on the Cityscapes autonomous driving dataset, which features a larger resolution closer to industrial practicality. As shown in Tab. 2, PL-DC continues to outperform under varied labeled data proportions. Specifically, compared with Supervised Mask2Former, our PL-DC improved mAP by $+16.4$, $+13.8$, $+8.7$, and $+8.3$ at 5%, 10%, 20%, and 30% labeled data, respectively, while GuidedDistillation still exhibited counterintuitive results at 5% and 10% labeled data. These results confirm that our PL-DC is robust and can be effectively generalized across different datasets.

Abalation Study

We conduct ablation studies on the proposed modules and hyper-parameters using the COCO dataset with 1% labeled data. A total of $73K$ iterations (**10 epochs**) are performed, which is shorter than the **50 epochs** used in the main experiments, allowing for faster validation during ablation.

Modules Validity. We ablate our proposed DF-ACAT, DICC, and PMUA, as depicted in Tab. 3. Removing DF-ACAT and replacing it with a coupled score threshold ($0.9 \times 0.85 = 0.765$) mainly diminishes the mAP, AP_m, and AP_l. This phenomenon occurs because, for medium and large objects, the competition between mask quality and class qual-

	mAP	AP_s	AP_m	AP_l
PL-DC (Ours)	**21.8**	**7.1**	**21.5**	**35.4**
- DF-ACAT	21.2 (↓ 0.6)	7.0 (↓ 0.1)	20.5 (↓ 1.0)	34.9 (↓ 0.5)
- DICC	20.5 (↓ 1.3)	6.1 (↓ 1.0)	20.4 (↓ 1.1)	35.0 (↓ 0.4)
- PMUA	20.6 (↓ 1.2)	6.3 (↓ 0.8)	20.1 (↓ 1.4)	34.6 (↓ 0.8)
- all above	20.6 (↓ 1.2)	6.4 (↓ 0.7)	20.3 (↓ 1.2)	34.7 (↓ 0.7)

Table 3: Ablation study (10 epochs) on COCO 1%. "-" means remove module. - DF-ACAT: remove DF-ACAT and replace it with a coupled score threshold (0.765) filtering. We evaluate the standard COCO metrics: mAP, AP_s for small, AP_m for medium, and AP_l for large objects.

$\mathcal{T}_c^{min}$	$\mathcal{T}_c^{max}$	$\mathcal{T}_m^{min}$	$\mathcal{T}_m^{max}$	mAP
–	–	–	–	21.1
0.6	0.8	0.7	0.85	21.4
0.6	0.8	0.8	0.85	21.5
0.6	0.8	0.8	0.9	21.6
0.7	0.8	0.8	0.9	21.6
0.7	0.85	0.8	0.9	**21.8**
0.7	0.85	0.9	0.9	21.7
0.8	0.85	0.9	0.9	21.4
0.9	0.9	0.9	0.95	21.0

Table 4: Different initial min&max thresholds in DF-ACAT.

ity prevents the coupled score threshold filtering mechanism from simultaneously evaluating both the class quality and mask quality of an instance effectively. Conversely, for small objects where the area is limited, class quality predominates in determining pseudo-label quality. Removing DICC results in a notable reduction in the AP_s and AP_m, likely due to their smaller visual features and higher susceptibility to classification errors. The removal of PMUA leads to a significant drop in the AP_m, attributable to the fact that the uncertainty area in medium objects represents a larger fraction of their total area. The combined removal of all modules results in a less marked decline in overall mAP than removing PMUA alone, suggesting a balanced compromise between object classification and mask segmentation capabilities.

Hyper-parameters Tuning. We abalate initial minimum and maximum thresholds in DF-ACAT in Tab. 4, EMA update rate α and the unsupervised loss $\mathcal{L}_{\text{unsup}}$ weight λ in Tab. 5. From Tab. 4, we derive three important observations: 1) When the initial threshold for mask quality is high, the model becomes relatively sensitive to the initial threshold for class quality. This is because, as depicted in Fig. 2 (b)(c), the mask quality does not reliably reflect the IoU alignment with the oracle ground truth, whereas class quality tends to correlate more accurately; 2) When the initial thresholds are set to moderate levels, DF-ACAT is very robust. This is because the initial threshold mainly acts on pseudo-label filtering in the early stage of training. As training progresses, the threshold will automatically converge to a range that is suitable for the model's capabilities; 3) when $\mathcal{T}_c^{min}$ and $\mathcal{T}_m^{min}$ are set excessively high (e.g., 0.9), pseudo-labels are over-filtered

EMA rate α		Weight λ for $\mathcal{L}_{\text{unsup}}$	
α	mAP	λ	mAP
0.5	19.9	0.5	21.2
0.7	20.0	1	**21.8**
0.9	20.2	2	18.7
0.99	20.5	4	10.0
0.999	21.2	8	6.6
0.9996	**21.8**	–	
0.9999	19.1	–	

Table 5: Ablation study on hyper-parameters in PL-DC.

in the early training phase, leading to poor recall and insufficient supervision, thereby degrading final performance. From Tab. 5 left part, it is evident that a smaller EMA rate α results in lower mAP, suggesting that the student model significantly influences the teacher model with each iteration, potentially propagating the negative effects of noisy pseudo-labels. Optimal performance is achieved at an EMA rate of 0.9996. However, further increasing α slows down updates to the teacher model, as it relies predominantly on its previous weights. As shown in Tab. 5 right part, the model performs optimally when the unsupervised loss weight λ is set to 1.0. Increasing this weight further leads to a sharp decline in mAP, indicating a detrimental effect on performance.

Qualitative Analysis

In Fig. I of the **Extended version**, we qualitatively analyze how each PL-DC component affects instance segmentation. Using $1k$ randomly sampled COCO *train2017* images, predictions are grouped into five types: **(Cor)** correct (IoU $>$ 0.5, correct class); **(Loc)** poor localization ($0 <$ IoU ≤ 0.5, correct class); **(Sim)** mislabeled as a semantically similar class; **(Oth)** mislabeled as an unrelated class; **(BG)** predicted on background (IoU = 0). From this analysis, we can draw four conclusions: (1) Removing DICC increases Oth and Sim errors, indicating its effectiveness in mitigating category confusion. (2) Removing PMUA raises the proportion of Loc and BG errors, suggesting it improves mask localization quality. (3) Removing DF-ACAT results in more BG, Sim, and Loc errors, showing its importance in regulating pseudo-label quality. (4) The most frequent failure mode involves confusion with other object categories, highlighting a potential for further improvement via advanced strategies.

Conclusion

We proposed PL-DC, a pseudo-label decoupling and correction framework for semi-supervised instance segmentation. PL-DC addresses instance-, category-, and pixel-level pseudo-label noise through three modules: Decoupled Filtering with Adaptive Class-Aware Thresholds, Dynamic Instance Category Correction, and Pixel-level Mask Uncertainty-Aware mask loss. Experiments on COCO and Cityscapes show that PL-DC delivers substantial gains and achieves state-of-the-art performance.

Acknowledgments

This work was supported by the National Science Fund for Distinguished Young Scholars (No.62025603 and No.62525605), National Natural Science Foundation of China (No. U21B2037, U22B2051, No. U23A20383, No. 62176222, No. 62176226, No. 62272401, No. 62576300).

References

Bai, M.; and Urtasun, R. 2017. Deep watershed transform for instance segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5221–5229.

Berrada, T.; Couprie, C.; Alahari, K.; and Verbeek, J. 2024. Guided Distillation for Semi-Supervised Instance Segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 475–483.

Berthelot, D.; Carlini, N.; Cubuk, E. D.; Kurakin, A.; Sohn, K.; Zhang, H.; and Raffel, C. 2019a. Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*.

Berthelot, D.; Carlini, N.; Goodfellow, I.; Papernot, N.; Oliver, A.; and Raffel, C. A. 2019b. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems (NIPS)*, 32.

Bolya, D.; Zhou, C.; Xiao, F.; and Lee, Y. J. 2019. Yolact: Real-time instance segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, 9157–9166.

Boykov, Y. Y.; and Jolly, M.-P. 2001. Interactive graph cuts for optimal boundary & region segmentation of objects in ND images. In *IEEE International Conference on Computer Vision (ICCV)*, volume 1, 105–112. IEEE.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, 213–229. Springer.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention Mask Transformer for Universal Image Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Cheng, B.; Schwing, A.; and Kirillov, A. 2021. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems (NIPS)*, 34: 17864–17875.

Comaniciu, D.; and Meer, P. 2002. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 24(5): 603–619.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3213–3223.

DeVries, T.; and Taylor, G. W. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.

Filipiak, D.; Zapała, A.; Tempczyk, P.; Fensel, A.; and Cygan, M. 2024. Polite teacher: Semi-supervised instance segmentation with mutual learning and pseudo-label thresholding. *IEEE Access*.

Gao, M.; Zhang, Z.; Yu, G.; Arık, S. Ö.; Davis, L. S.; and Pfister, T. 2020. Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In *European Conference on Computer Vision (ECCV)*, 510–526. Springer.

Gao, N.; Shan, Y.; Wang, Y.; Zhao, X.; Yu, Y.; Yang, M.; and Huang, K. 2019. Ssap: Single-shot instance segmentation with affinity pyramid. In *IEEE International Conference on Computer Vision (ICCV)*, 642–651.

Girshick, R. 2015. Fast r-cnn. In *IEEE International Conference on Computer Vision (ICCV)*, 1440–1448.

Gupta, A.; Dollár, P.; and Girshick, R. 2019. LVIS: A Dataset for Large Vocabulary Instance Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *IEEE International Conference on Computer Vision (ICCV)*, 2961–2969.

Hu, J.; Chen, C.; Cao, L.; Zhang, S.; Shu, A.; Jiang, G.; and Ji, R. 2023. Pseudo-label Alignment for Semi-supervised Instance Segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, 16291–16301.

Jeong, J.; Lee, S.; Kim, J.; and Kwak, N. 2019a. Consistency-based semi-supervised learning for object detection. *Advances in Neural Information Processing Systems (NIPS)*, 32.

Jeong, J.; Lee, S.; Kim, J.; and Kwak, N. 2019b. Consistency-based Semi-supervised Learning for Object detection. *Advances in Neural Information Processing Systems (NIPS)*.

Lee, Y.; and Park, J. 2020. Centermask: Real-time anchor-free instance segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13906–13915.

Li, Y.; Mao, H.; Girshick, R.; and He, K. 2022. Exploring plain vision transformer backbones for object detection. In *European Conference on Computer Vision (ECCV)*, 280–296. Springer.

Lin, J.; Shen, Y.; Wang, B.; Lin, S.; Li, K.; and Cao, L. 2024. Weakly supervised open-vocabulary object detection. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, 3404–3412.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft Coco: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*.

Liu, S.; Jia, J.; Fidler, S.; and Urtasun, R. 2017. Sgn: Sequential grouping networks for instance segmentation. In *IEEE International Conference on Computer Vision (ICCV)*, 3496–3504.

Liu, Y.-C.; Ma, C.-Y.; He, Z.; Kuo, C.-W.; Chen, K.; Zhang, P.; Wu, B.; Kira, Z.; and Vajda, P. 2021. Unbiased teacher for

semi-supervised object detection. *International Conference on Learning Representations (ICLR)*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Luo, W.; and Yang, M. 2020. Semi-supervised Semantic Segmentation via Strong-Weak Dual-Branch Network. *Springer International Publishing eBooks*.

Mi, P.; Lin, J.; Zhou, Y.; Shen, Y.; Luo, G.; Sun, X.; Cao, L.; Fu, R.; Xu, Q.; and Ji, R. 2022. Active teacher for semi-supervised object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14482–14491.

Ouali, Y.; Hudelot, C.; and Tami, M. 2020. Semi-Supervised Semantic Segmentation with Cross-Consistency Training. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 8748–8763. PMLR.

Radosavovic, I.; Dollár, P.; Girshick, R. B.; Gkioxari, G.; and He, K. 2018. Data Distillation: Towards Omni-Supervised Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sohn, K.; Zhang, Z.; Li, C.-L.; Zhang, H.; Lee, C.-Y.; and Pfister, T. 2020. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*.

Tarvainen, A.; and Valpola, H. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems (NIPS)*, 30.

Wang, Z.; Li, Y.; and Wang, S. 2022. Noisy Boundaries: Lemon or Lemonade for Semi-supervised Instance Segmentation? *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16805–16814.

Wu, Y.; Kirillov, A.; Massa, F.; Lo, W.-Y.; and Girshick, R. 2019. Detectron2. https://github.com/facebookresearch/detectron2.

Xu, M.; Zhang, Z.; Hu, H.; Wang, J.; Wang, L.; Wei, F.; Bai, X.; and Liu, Z. 2021. End-to-end semi-supervised object detection with soft teacher. In *IEEE International Conference on Computer Vision (ICCV)*, 3060–3069.

Yu, F.; Chen, H.; Wang, X.; Xian, W.; Chen, Y.; Liu, F.; Madhavan, V.; and Darrell, T. 2020. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2636–2645.