Pseudo-label Alignment for Semi-supervised Instance Segmentation

Jie Hu[1†], Chen Chen[1†], Liujuan Cao[1*], Shengchuan Zhang[1], Annan Shu[2],
Guannan Jiang[2], and Rongrong Ji[1]
[1]Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University
[2]Contemporary Amperex Technology Co. Limited

Abstract

Pseudo-labeling is significant for semi-supervised instance segmentation, which generates instance masks and classes from unannotated images for subsequent training. However, in existing pipelines, pseudo-labels that contain valuable information may be directly filtered out due to mismatches in class and mask quality. To address this issue, we propose a novel framework, called pseudo-label aligning instance segmentation (PAIS), in this paper. In PAIS, we devise a dynamic aligning loss (DALoss) that adjusts the weights of semi-supervised loss terms with varying class and mask score pairs. Through extensive experiments conducted on the COCO and Cityscapes datasets, we demonstrate that PAIS is a promising framework for semi-supervised instance segmentation, particularly in cases where labeled data is severely limited. Notably, with just 1% labeled data, PAIS achieves 21.2 mAP (based on Mask-RCNN) and 19.9 mAP (based on K-Net) on the COCO dataset, outperforming the current state-of-the-art model, i.e.*, NoisyBoundary with 7.7 mAP, by a margin of over 12 points. Code is available at:* https://github.com/hujiecpp/PAIS.

1. Introduction

Semi-supervised instance segmentation aims to alleviate the significant burden of human labeling by utilizing a small amount of labeled data in conjunction with abundant unlabeled data [35, 47, 50]. Existing semi-supervised instance segmentation pipelines typically generate pseudo-labels from unlabeled images, which are then used to train the models together with labeled images. Therefore, pseudo-labels play a crucial role in semi-supervised instance segmentation. The generation of pseudo-masks, pseudo-classes, and pseudo-boxes from unlabeled images improves the model training. However, current semi-supervised instance segmentation frameworks do not fully leverage the potential of such pseudo-labels. Specifically, pseudo-labels with mismatched class and mask scores are often filtered out by fixed thresholds, leading to the exclusion of valuable information that could aid the model training. For instance, pseudo-labels with high-quality masks but low class scores would be filtered out by a class threshold, resulting in the loss of the pixel-level information.

In this paper, we present a new semi-supervised framework, termed pseudo-label aligning for instance segmentation (PAIS), aiming to improve the utilization of filtered pseudo-labels. As illustrated in Fig. 1, the main challenge of PAIS lies in the mismatched scores between pseudo-classes and pseudo-masks. The classification score and mask intersection over union (IoU) are misaligned in assessing the quality of pseudo-labels. As a result, masks with high IoUs would be filtered out due to low classification scores, and vice versa. In the mean time, lowering the threshold of both scores introduces incorrect classes or low-quality masks into the semi-supervised training. To overcome this dilemma, we propose a dynamic aligning loss (DALoss) that softly re-weights the classification and the segmentation losses based on the quality of different pseudo-labels. Specifically, DALoss penalizes low-score pseudo-labels rather than filtering them and promotes high-score ones, to adjust their contribution to the final loss function. Our experiments on the COCO and Cityscape datasets demonstrate the effectiveness of the proposed PAIS framework. Specifically, with only 1% labeled data, PAIS achieves 21.2 mAP and 19.9 mAP on the COCO dataset using Mask-RCNN [13] and K-Net [53], respectively. This outperforms the current state-of-the-art model, NoisyBoundary [47], by more than 12 points.

Our contributions can be summarized as follows:

- We propose a novel pseudo-label aligning framework for semi-supervised instance segmentation, called PAIS, which unleashes the potential of utilizing pixel-level pseudo-labels in semi-supervised instance segmentation. Furthermore, to the best of our knowledge,

[*]Corresponding author. [†]Equal contribution.

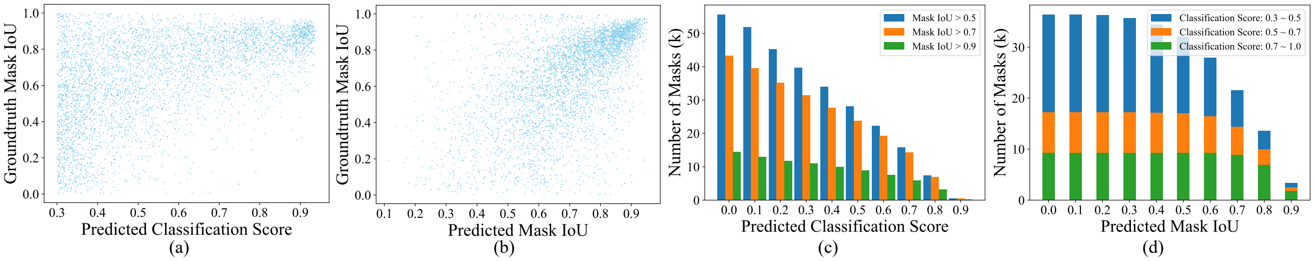

Figure 1. **Pseudo-label wasting in semi-supervised instance segmentation.** To illustrate this issue, we employed the K-Net model [53] trained on 10% labeled images and randomly sampled 5k unlabeled images, from the COCO `train2017` dataset, to plot the figures. In (a), we present the predicted classification score w.r.t. the ground truth mask intersection over union (IoU), revealing that the predicted classification score is inadequate in evaluating mask quality. In (b), we added a mask IoU prediction branch to the model and plotted the predicted mask IoU w.r.t. the ground truth mask IoU, demonstrating that the predicted mask IoU can represent the mask quality. However, as shown in (c), a significant number of pseudo-labels with high-quality masks, *i.e*., mask IoU $>$ 0.7, will be filtered out with high classification thresholds. Additionally, (d) shows that pseudo-labels with high-quality masks can also have correspondingly low classification scores. Therefore, the misalignment between the mask and class scores leads to a significant number of valuable pseudo-labels being excluded during semi-supervised training.

PAIS is the first framework that can be adapt to box-free instance segmentation models.

- We introduce a new loss function, named dynamic aligning loss (DALoss), which incorporates pseudo-labels with diverse class and mask qualities into the semi-supervised training process. DALoss consistently enhances the performance of box-free and box-dependent instance segmentation frameworks.
- We conduct comprehensive experiments on the COCO and Cityscapes datasets to evaluate PAIS. In particular, PAIS achieves state-of-the-art results on the COCO dataset, *i.e*., 19.9, 27.6, and 31.1 mAP for the box-free pipeline K-Net [53], and 21.2, 29.3, and 31.1 mAP for the box-dependent pipeline Mask-RCNN [13], with 1%, 5%, and 10% labeled data, respectively.

2. Related Work

Semi-supervised Image Classification. In image classification, semi-supervised learning has been extensively explored, and the methods can be classified into two categories: pseudo-label-based and consistency-regularization-based methods. Specifically, pseudo-label-based methods [40, 21] leverage pre-trained models to generate annotations for the unlabeled images to train the model. In contrast, consistency-regularization-based methods [1, 36, 20, 39, 48, 2, 7, 32, 31, 23] incorporate various data augmentation techniques such as random regularization [10] and adversarial perturbation [33] to generate different inputs for one image and enforce consistency between these inputs during training. FixMatch [41] combines the consistency-regularization-based techniques with a pseudo-label-based framework by applying a strong-weak data augmentation pipeline to input images and enforcing consistency between the augmented images. In this work, we follow the pseudo-label-based methods and also use strong-weak data augmentation during training in PAIS.

Semi-supervised Object Detection. From the very beginning, STAC [42] proposed the use of pseudo-labels and consistency training for semi-supervised object detection. However, the effectiveness of the method was limited by the two-stage training pipeline similar to that of Noisy Student [49], where the pseudo-labels were generated from pre-trained models and were not updated along with the model training. After STAC, several studies [50, 54, 43, 51, 28] incorporated the idea of exponential moving average (EMA) from MeanTeacher [44]. The teacher model and pseudo-labels are updated after each training iteration to generate instant pseudo-labels, making the entire pipeline end-to-end trainable. Additionally, Unbiased Teacher [28] utilized Focal loss [24] instead of traditional cross-entropy loss to alleviate the problem of unbalanced pseudo-labels. In this paper, we also follow the idea of incorporating EMA into the proposed PAIS framework, with a focus on integrating pixel-level annotations into the training process.

Fully-supervised Instance Segmentation. Instance segmentation aims to provide pixel-level predictions for each object instance in an image. Existing methods can be classified into three categories: top-down (or box-dependent) methods, bottom-up methods, and direct segmentation (or box-free) methods. Top-down methods [15, 27, 38, 37, 29] such as Mask R-CNN [13], YOLACT [3], and CenterMask [22] generate bounding boxes first and then segment the objects within the boxes. Bottom-up methods [52, 8, 26, 11, 16], regard instance segmentation as a label-then-cluster problem, which classify each pixel first and then group the pixels into an arbitrary number of object instances. Direct segmentation or box-free methods such as SOLO [45, 46], K-Net [53], MaskFormer [5, 4] and

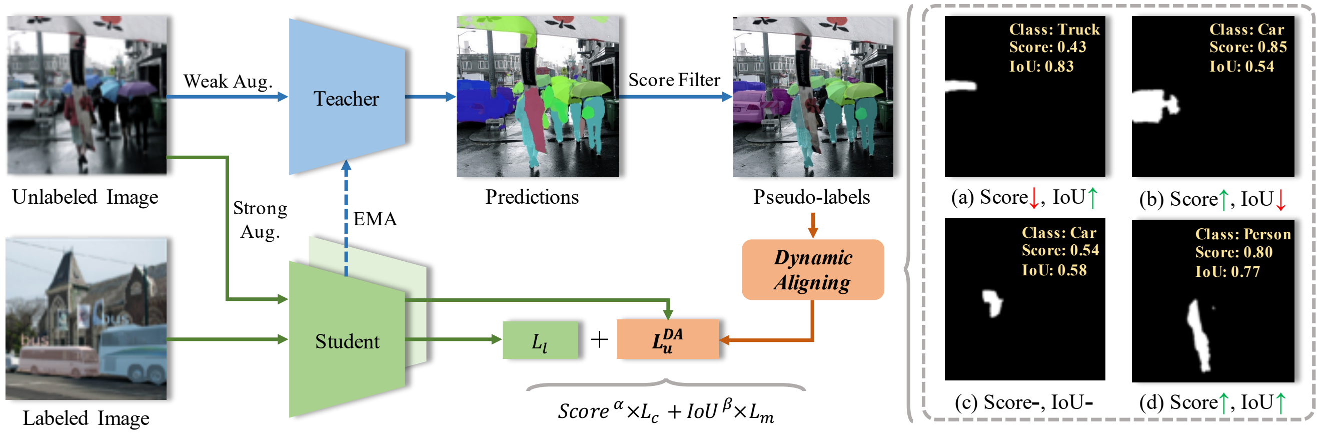

Figure 2. **Framework of pseudo-label aligning for semi-supervised instance segmentation (PAIS).** The focus of PAIS is to explore the utility of pixel-level pseudo-labels for instance segmentation. To this end, we propose a novel approach called dynamic aligning loss (DALoss), which enables the incorporation of pseudo-labels with varying quality during training. This is particularly significant when the number of labeled images is limited. With DALoss, a significant number of predictions from the teacher model can be leveraged in the semi-supervised training process, rather than being discarded. Specifically, we consider three types of pseudo-masks with varying quality: (a) pseudo-masks with incorrect classes but high-quality masks, *i.e.*, low classification scores but high mask IoUs; (b) pseudo-masks with accurate classes but low-quality masks, *i.e.*, high classification scores but low mask IoUs; and (c) pseudo-masks with medium qualities.

SOTR [12] deal with instance segmentation without bounding box detection. Any of the aforementioned instance segmentation methods can be implemented into PAIS, and in this work, we present two examples, one from the box-dependent category, Mask R-CNN, and another from the box-free category, K-Net.

Semi-supervised Instance Segmentation. Semi-supervised instance segmentation is commonly considered to be a sub-task of semi-supervised object detection [50, 28]. Consequently, existing frameworks rely heavily on bounding boxes, in which segmentation performance is strongly dependent on detection performance. Among them, NoisyBoundary [47] was the first to formally propose the semi-supervised instance segmentation task. Recent efforts have been made to construct box-free pipelines for fully-supervised instance segmentation [45, 46, 53, 5]. In this paper, we investigate PAIS in both box-free and box-dependent instance segmentation, revealing the potential of fully utilizing pixel-level annotations. In contrast to the recently-proposed PoliteTeacher [9], which filters out pseudo-labels with low confidence, the proposed PAIS leverages them. It is a novel and effective way of utilizing noisy pseudo-labels for semi-supervised learning.

3. Method

3.1. Task Formulation

The goal of PAIS is to better leverage unlabeled images with a limited number of pixel-labeled images to boost the performance of semi-supervised instance segmentation. The PAIS framework consists of three key steps: (1) pseudo-label generation, (2) dynamic pseudo-label alignment, and (3) end-to-end model training. In the pseudo-label generation step, we introduce a mask scoring branch [17] that predicts mask IoUs as an additional metric along with classification scores to assess the quality of pseudo-labels. In the dynamic aligning step, we re-weight the loss terms based on the quality of different pseudo-labels. Finally, the teacher and student models are trained using exponential moving average (EMA) [44]. The overall framework of PAIS is illustrated in Fig. 2, and we introduce the key steps in detail as follows. It is important to note that the PAIS framework can be applied to any box-dependent or box-free framework for enhanced exploitation of pseudo-labels in semi-supervised instance segmentation. We instantiate two examples by Mask-RCNN [13] and K-Net [53] in this paper, but do not restrict to them.

3.2. Pseudo Label Generation

In the pseudo-label generation, we apply weak data augmentation to the unlabeled images and input them into a teacher model. The weak data augmentation includes scaling, horizontal flipping, and other augmentation operations that do not alter the image's content. The teacher model produces a set of pseudo-labels, including masks, boxes, and classification scores for each input image. In box-free pipelines, the box predictions are optional.

As depicted in Fig. 1(a)(c), using the classification score alone is inadequate to measure the quality of predicted masks. To address this, we incorporate a mask scoring branch into the pipeline to predict mask Intersection over Union (IoU) for evaluating mask quality. As shown in Fig. 1(b), the predicted mask IoUs can be used to measure mask quality effectively, which has also been verified

Algorithm 1 Pseudo-label Alignment

```
Initialize teacher and student models randomly.
Repeat
    Apply weak augmentation to unlabeled images.
    Obtain pseudo-labels $\{\boldsymbol{m}_i, \boldsymbol{b}_i, \boldsymbol{c}_i, s_i | i = 1, \ldots, N\}$ from teacher model via thresholds $\tau_{cls}, \tau_{iou}$.
    Apply strong augmentation to unlabeled images.
    Obtain predictions $\widetilde{\boldsymbol{m}}, \widetilde{\boldsymbol{b}}, \widetilde{\boldsymbol{c}}$ for unlabeled images.
    Calculate $\mathcal{L}_u$ for unlabeled images by Eq. 5.
    Inference labeled images by student model.
    Calculate $\mathcal{L}_l$ for labeled images.
    Train student model with the loss $\mathcal{L}_{semi}$ in Eq. 1.
    Update teacher model via EMA.
Until scheduled epochs.
```

in MS-RCNN [17]. After generating pseudo-labels in the previous step, the predictions are filtered using two thresholds: a classification threshold τ_{cls} and a mask IoU threshold τ_{iou}. The resulting set of pseudo-labels includes N elements, each containing a mask $\boldsymbol{m}_i \in \mathbb{R}^{H \times W}$, a bounding box $\boldsymbol{b}_i \in \mathbb{R}^4$, a classification score $\boldsymbol{c}_i \in \mathbb{R}^L$ (with an additional background category), and a mask IoU score $s_i \in \mathbb{R}$. Note that H, W denote the mask resolution, and L is the number of classes.

3.3. Dynamic Aligning Loss

Although good pseudo-labels can be obtained by setting high thresholds for classification scores and mask IoUs, a large number of predictions with misaligned mask and classification qualities are discarded (as illustrated in Fig.1(c)(d)). These misaligned predictions can be useful since the amount of labeled data is limited in the semi-supervised setting. Fig.2 shows examples of pseudo-labels that can be included in the training process. If the misaligned pseudo-labels are directly used in semi-supervised learning, the incorrect predictions on classes or masks can introduce significant noise. To reduce the noise, we propose a dynamic aligning loss (DALoss).

Vanilla Loss for PAIS. In semi-supervised instance segmentation, the loss function can be decomposed into two terms for labeled and unlabeled images, as:

$$\mathcal{L}_{semi} = \lambda_l \mathcal{L}_l + \lambda_u \mathcal{L}_u, \tag{1}$$

where λ_l and λ_u are the hyper-parameters for balancing the loss terms. The loss for labeled images can be defined using the functions commonly employed in instance segmentation, augmented with an additional binary cross-entropy term for regressing mask IoUs. Given the pseudo-labels from the teacher model and the predictions from the student model, we formulate the loss for unlabeled images as:

$$\begin{aligned} \mathcal{L}_u = & \frac{1}{N_{pos}} \sum_{i=1}^{N_{pos}} \big(\lambda_b \mathcal{L}_b(\widetilde{\boldsymbol{b}}_{\nu(i)}, \boldsymbol{b}_{\mu(i)}) \\ & + \lambda_c \mathcal{L}_c(\widetilde{\boldsymbol{c}}_{\nu(i)}, \boldsymbol{c}^o_{\mu(i)}) + \lambda_m \mathcal{L}_m(\widetilde{\boldsymbol{m}}_{\nu(i)}, \boldsymbol{m}^b_{\mu(i)})\big) \\ & + \frac{1}{N_{neg}} \sum_{i=1}^{N_{neg}} \big(\lambda_c \mathcal{L}_c(\widetilde{\boldsymbol{c}}_i, \boldsymbol{c}^o_{neg})\big), \end{aligned} \tag{2}$$

where $\mathcal{L}_b$ includes the box IoU loss and the L1 loss, $\mathcal{L}_c$ denotes the cross-entropy loss, $\mathcal{L}_m$ is the dice loss. In Eq. 2, $\widetilde{\boldsymbol{c}}_i, \widetilde{\boldsymbol{b}}_i, \widetilde{\boldsymbol{m}}_i$ represent predictions from the student model. The pseudo-scores $\boldsymbol{c}_{\mu(i)}$ for classification are converted to one-hot vectors $\boldsymbol{c}^o_{\mu(i)}$ for the category with the highest score. The pseudo-masks $\boldsymbol{m}_{\mu(i)}$ are activated by sigmoid function and discretized into binary value as $\boldsymbol{m}^b_{\mu(i)}$. The one-hot vector for the background class is denoted by $\boldsymbol{c}^o_{neg}$. The label indexing functions, $\nu(i)$ and $\mu(i)$, match predictions with pseudo-labels for training, which are defined in terms of different pseudo-label assigning strategies. In our implementations, the one-to-many assignment defines the label indexing functions $\nu(i) = i$ and $\mu(i)$ as:

$$\mu(i) = \arg\min_{t \in [1,N]} \big(\mathcal{L}_b(\widetilde{\boldsymbol{b}}_i, \boldsymbol{b}_t)\big). \tag{3}$$

Instead, the one-to-one assignment defines the label indicting functions $\mu(i) = i$, and finds the optimal $\nu^*(\cdot)$ via:

$$\begin{aligned} \nu^* = \arg\min_{\nu} & \sum_{i=1}^{t} \big(\lambda_b \mathcal{L}_b(\widetilde{\boldsymbol{b}}_{\nu(i)}, \boldsymbol{b}_i) + \lambda_c \mathcal{L}_c(\widetilde{\boldsymbol{c}}_{\nu(i)}, \boldsymbol{c}^o_i) \\ & + \lambda_m \mathcal{L}_m(\widetilde{\boldsymbol{m}}_{\nu(i)}, \boldsymbol{m}^b_i) - s_i\big). \end{aligned} \tag{4}$$

Note that the loss terms for regressing boxes are optional in box-free instance segmentation frameworks.

Dynamic Aligning Loss for PAIS. To optimize the model using pixel-level pseudo-labels, we propose to replace Eq. 2 with DALoss. Since the teacher model provides ideal metrics for measuring the quality of classification scores and mask IoUs, we propose to adjust the weight of loss terms based on the quality of the pseudo-labels. This is achieved by using the following equation:

$$\begin{aligned} \mathcal{L}^{DA}_u = & \frac{1}{N_{pos}} \sum_{i=1}^{N_{pos}} \big(\lambda_b \mathcal{L}_b(\widetilde{\boldsymbol{b}}_{\nu(i)}, \boldsymbol{b}_{\mu(i)}) \\ & + \lambda_c (c^h_{\mu(i)})^{\alpha} \mathcal{L}_c(\widetilde{\boldsymbol{c}}_{\nu(i)}, \boldsymbol{c}^o_{\mu(i)}) \\ & + \lambda_m (s_{\mu(i)})^{\beta} \mathcal{L}_m(\widetilde{\boldsymbol{m}}_{\nu(i)}, \boldsymbol{m}^b_{\mu(i)})\big) \\ & + \frac{1}{N_{neg}} \sum_{i=1}^{N_{neg}} \big(\lambda_c \mathcal{L}_c(\widetilde{\boldsymbol{c}}_i, \boldsymbol{c}^o_{neg})\big), \end{aligned} \tag{5}$$

where $c^h_{\mu(i)}$ denotes the highest classification score, and $s_{\mu(i)}$ denotes the mask IoU from the $\mu(i)$-th pseudo-label,

Method	1%	5%	10%	100%
Mask-RCNN [13], supervised*	3.5	17.3	22.0	34.5
Mask-RCNN† [13], supervised*	3.5	17.4	21.9	37.1
DD [35]	3.8	20.4	24.2	35.7
Noisy Boundaries [47]	7.7	24.9	29.2	38.6
PAIS, on Mask-RCNN, *ours*	**21.2**	**29.3**	**31.1**	**39.5**

Table 1. **Comparison to state-of-the-art** semi-supervised instance segmentation methods on the COCO `val2017`. † denotes using the same data augmentation as semi-supervised training. * denotes data from NoisyBoundary [47].

Method	5%	10%	20%	30%
Mask-RCNN [13], supervised*	11.8	16.8	22.3	26.3
Mask-RCNN† [13], supervised*	11.3	16.4	22.6	26.6
DD [35]	13.7	19.2	24.6	27.4
STAC [42]	11.9	18.2	22.9	29.0
CSD [18]	14.1	17.9	24.6	27.5
CCT [34]	15.2	18.6	24.7	26.5
Dual-branch [30]	13.9	18.9	24.0	28.9
Ubteacher [28]	16.0	20.0	27.1	28.0
Noisy Boundaries [47]	17.1	22.1	29.0	32.4
PAIS, on Mask-RCNN, *ours*	**18.0**	**22.9**	**29.2**	**32.8**

Table 2. **Comparison to state-of-the-art** semi-supervised instance segmentation methods on the Cityscapes validation set. † denotes using the same data augmentation as semi-supervised training. * denotes data from NoisyBoundary [47].

α, β are the hyper-parameters. Specifically, Eq. 5 adjusts the weights for the pseudo-labels conditioned on their qualities, *i.e.*, dynamically dependent on the input images. For instance, for a pseudo-label with a low classification score and high mask IoU, DALoss encourages the segmentation loss while constraining the classification loss.

3.4. End-to-End Model Training

Inspired by [44], we employ EMA with the strong-weak data augmentation for PAIS. Specifically, unlabeled images undergo both strong and weak data augmentations and are then fed into the student and teacher models, respectively. The student model is trained to produce consistent results with the pseudo-labels, and the teacher model is updated by EMA. The training pipeline for PAIS is presented in Alg. 1.

4. Experiments

4.1. Datasets and Evaluation Metrics

We conducted extensive experiments on the COCO [25] and Cityscapes [6] datasets to study the proposed PAIS. The COCO dataset consists of 118k images with 80-class instance labels, as well as 123k unlabeled images. The Cityscapes dataset contains urban street-view scenes and has 8 instance categories in 2.9k training images and 0.5k validation images. For the COCO dataset, we randomly sampled 1%, 5%, and 10% of the images from the `train2017` split as labeled data and treated the rest as unlabeled data following common settings. In addition, we also used the full COCO `train2017` as labeled data and incorporated the 123k unlabeled data from COCO `unlabel2017` to train the PAIS models. For the Cityscapes dataset, we randomly sampled 5%, 10%, 20%, and 30% of the images from the training set as labeled data and treated the remaining as unlabeled data following the common settings. We evaluated the PAIS models on the validation sets of the COCO and Cityscapes datasets, and reported standard COCO metrics including AP, AP50, AP75 (averaged over IoU thresholds), and AP$_S$, AP$_M$, AP$_L$ (AP for instances of different scales).

4.2. Implementation Details

We provide two examples of implementing PAIS with K-Net [53] and Mask-RCNN [13]. The models are trained using AdamW with a learning rate of 0.0001 for K-Net, and SGD with a learning rate of 0.01 for Mask-RCNN. The hyper-parameters λ_l and λ_u, which balance the loss terms for labeled and unlabeled images, are set to 1.0 and 0.3 for K-Net, and 1.0 and 1.5 for Mask-RCNN. We set the thresholds τ_{cls} and τ_{iou} experimentally as 0.35 and 0.30, respectively. For bipartite matching loss, we use the same hyperparameters as in [53]. The loss balancing parameters for box, class, and mask are set as λ_b=2.0, λ_c=4.0, and λ_m=1.0, respectively. We train the models on 4 GPUs with 4 images per GPU (1 labeled and 3 unlabeled images) for 220k iterations, unless otherwise specified. The teacher model is updated via EMA with a momentum of 0.999. We use ResNet50 [14] as the backbone for these models.

4.3. Main Results

Comparison to state-of-the-art semi-supervised instance segmentation frameworks. In Tab. 1, we compare the performance of models trained with PAIS to the state-of-the-art semi-supervised instance segmentation frameworks on the COCO dataset. The results demonstrate that both K-Net and Mask-RCNN trained with PAIS surpass the previous methods DD [35] and Noisy Boundaries [47] by a large margin, especially when the number of labeled data is very limited (with only 1% or 5% labeled images). Specifically, when using 1% labeled COCO images, PAIS with K-Net achieves 19.9 mask mAP, which is 12.2 points higher than Noisy Boundaries. Interestingly, the proposed PAIS brings about better performance for Mask-RCNN when the percentage of labeled data is 1% or 5%, even though it originally has a inferior performance in fully-supervised instance segmentation compared to K-Net. To explain, the bounding boxes may provide better optimization for Mask-RCNN models when the labeled data is limited. Finally, when using 10% and 100% labeled images, PAIS with K-Net out-

Method	1% COCO			5% COCO			10% COCO		
	AP	AP$_{0.5}$	AP$_{0.75}$	AP	AP$_{0.5}$	AP$_{0.75}$	AP	AP$_{0.5}$	AP$_{0.75}$
Box-free Instance Segmentation									
K-Net [53], supervised	8.03±0.25	16.33±0.31	7.00±0.25	17.4±0.22	30.08±0.28	16.83±0.21	21.63±0.21	37.43±0.38	21.8±0.20
K-Net† [53], supervised	11.63±0.05	22.30±0.08	10.95±0.10	22.28±0.05	38.54±0.09	22.7±0.07	26.53±0.12	44.87±0.15	27.20±0.17
K-Net, PAIS w/o DALoss	17.77±0.06	32.17±0.15	17.53±0.12	25.40±0.08	43.12±0.05	26.05±0.13	29.30±0.07	48.64±0.05	30.53±0.13
K-Net, PAIS	**19.78**±0.10	**35.48**±0.15	**19.65**±0.06	**27.53**±0.06	**45.63**±0.06	**28.70**±0.10	**31.04**±0.06	**50.50**±0.07	**32.34**±0.05
Box-dependent Instance Segmentation									
Mask-RCNN [13], supervised*	3.5	-	-	17.3	-	-	22.0	-	-
Mask-RCNN† [13], supervised	11.54±0.09	19.86±0.11	11.64±0.09	22.35±0.06	37.98±0.10	23.14±0.11	27.07±0.06	45.10±0.10	28.67±0.06
Mask-RCNN, PAIS w/o DALoss	20.13±0.06	33.23±0.15	21.27±0.06	27.36±0.06	44.10±0.10	29.27±0.06	29.77±0.06	47.70±0.10	31.97±0.06
Mask-RCNN, PAIS	**21.12**±0.05	**36.03**±0.05	**22.75**±0.10	**29.28**±0.13	**47.25**±0.13	**31.20**±0.22	**31.03**±0.06	**49.83**±0.12	**33.23**±0.06

Table 3. **Results of PAIS with extremely limited number of labeled images.** We randomly sampled 1%, 5%, and 10% of the labeled images from the COCO `train2017` dataset, and repeated each experiment three times. The average values with standard deviation are reported. † denotes using the same data augmentation as semi-supervised training. * denotes data from NoisyBoundary [47]. We can see that PAIS achieves good performance on both box-free and box-independent instance segmentation frameworks.

Method	AP	AP$_{0.5}$	AP$_{0.75}$	AP$_S$	AP$_M$	AP$_L$
Box-free Instance Segmentation						
K-Net [53], supervised	37.8	60.3	39.9	16.9	41.2	57.5
K-Net† [53], supervised	38.4	61.4	40.3	17.6	41.8	58.0
K-Net, PAIS w/o DALoss	39.4	62.2	41.6	18.5	42.8	59.2
K-Net, PAIS	**40.8**	**63.5**	**43.3**	**19.2**	**44.4**	**61.4**
Box-dependent Instance Segmentation						
Mask-RCNN [13], supervised	37.1	58.5	39.7	18.7	39.6	53.9
Mask-RCNN† [13], supervised	37.5	58.9	40.4	18.6	40.2	53.8
Mask-RCNN, PAIS w/o DALoss	38.4	59.7	41.5	19.4	41.1	55.0
Mask-RCNN, PAIS	**39.5**	**60.6**	**43.0**	**19.9**	**42.4**	**56.6**

Table 4. **Results of PAIS with abundant labeled images.** We utilize all the labeled images in the COCO `train2017` dataset, and supplement them with the unlabeled images in the COCO `unlabel2017` dataset for semi-supervised training. † denotes using the same data augmentation as semi-supervised training.

performs PAIS with Mask-RCNN. In Tab. 2, we compare PAIS with state-of-the-art methods on the Cityscape dataset, in which PAIS also achieves better performance than the predominant models. Specifically, we report NoisyBoundaries [47] w/o FocalLoss in Tab. 2, as we do not apply FocalLoss in PAIS. This ensures a fair and consistent comparison. Furthermore, we also add FocalLoss to PAIS on 10% Cityscapes, which achieves 25.1%, surpassing 23.7% of NoisyBoundaries w/ FocalLoss.

The more significant performance improvement on the COCO dataset validates the effectiveness of our method for solving the noisy pseudo-label problem. The COCO dataset has 80 instance categories, while the Cityspaces dataset only has 8 instance categories. This implies that the COCO dataset can provide more diverse and informative pseudo-labels, which matches our goal to utilize noisy pseudo-labels for semi-supervised learning.

Results with an extremely limited number of labeled images. To demonstrate the effectiveness of PAIS, we conduct experiments with extremely limited numbers of labeled images, as shown in Table 3. Specifically, we compare the performance of various models trained with randomly sampled 1%, 5%, and 10% labeled images. First, we train supervised models, K-Net (supervised) and Mask-RCNN (supervised), with the limited labeled images. Second, we train the same supervised models with the same data augmentation used in the semi-supervised setting, denoted as K-Net† (supervised) and Mask-RCNN† (supervised). Third, we train PAIS models without DALoss, denoted as PAIS w/o DALoss on K-Net and Mask-RCNN. Lastly, we train the PAIS models with DALoss, denoted as PAIS on K-Net and Mask-RCNN. All models are trained three times, and the reported results are averaged.

Based on the results of K-Net (supervised) and Mask-RCNN (supervised), it can be observed that fully-supervised models perform poorly when the number of labeled images is limited. The results of K-Net† (supervised) and Mask-RCNN† (supervised) show slight improvement with weak data augmentation from the semi-supervised setting. Interestingly, while K-Net outperforms Mask-RCNN in the fully-supervised setting, their performance is similar when the number of labeled images is limited, as indicated in the table. Comparing the performance of K-Net (PAIS w/o DALoss) and Mask-RCNN (PAIS w/o DALoss) with that of the supervised setting reveals significant improvement. By introducing the dynamically re-weighting process via DALoss, the performance is further improved. For instance, with only 1% labeled images, the mAP of K-Net and Mask-RCNN improved from 11.63 and 11.54 to 19.78 and 21.12, respectively, resulting in approximately +8.15 and +9.58 points improvement to the performance. Additionally, the improvement over AP$_{0.5}$ and AP$_{0.75}$ suggests that DALoss considers masks with moderate quality during training, which further benefits semi-supervised learning.

Results with abundant labeled images. In Tab. 4, we investigate the performance of semi-supervised learning

Cls.	IoU.	Mask.	AP	$\text{AP}_{0.5}$	$\text{AP}_{0.75}$	AP_S	AP_M	AP_L
			29.3	48.7	30.5	11.3	31.4	45.5
✓			29.7	49.0	31.0	11.4	31.8	46.3
	✓		29.4	48.9	30.6	11.7	31.5	46.0
	✓	✓	30.4	50.1	31.9	11.9	32.8	47.6
✓	✓	✓	**31.1**	**50.6**	**32.4**	**12.3**	**33.3**	**48.3**

Table 5. **Effectiveness of different loss terms in DALoss**. *Cls.* denotes the terms of dynamic aligned classification scores. *IoU.* denotes adding the mask IoU branch to the model. *Mask.* denotes using the terms of dynamic aligned mask scores.

α, β	AP	$\text{AP}_{0.5}$	$\text{AP}_{0.75}$	AP_S	AP_M	AP_L
1	29.9	49.1	31.0	10.9	32.1	47.0
2	30.2	49.5	31.1	11.1	32.2	47.1
3	30.5	50.1	31.7	11.6	32.9	47.6
4	**31.1**	**50.6**	**32.4**	**12.3**	**33.3**	**48.3**

Table 6. **Hyper-parameters in DALoss.** Considering that the classification and segmentation are both important, we simply set $\alpha = \beta$ to investigate the influence of hyper-parameters.

when abundant labeled data is available. Specifically, we use the entire COCO `train2017` dataset as labeled data and COCO `unlabel2017` dataset as unlabeled data to train the models. The results show that the semi-supervised learning approach also leads to a performance gain. For instance, K-Net (PAIS) achieves a performance gain of approximately 1.0 point on mAP from semi-supervised learning. Additionally, the proposed DALoss consistently improves the performance of both box-dependent and box-free instance segmentation frameworks.

4.4. Ablation Study

In our ablation study, we investigate several aspects that can impact the performance of PAIS, including the utilization of various loss terms, the setting of hyper-parameters, the threshold values, the varying ratios of labeled and unlabeled images, and the convergence times. We perform the ablation studies on the COCO dataset using K-Net and Mask-RCNN, which are trained via PAIS under the setting of 10% labeled images.

Effectiveness of different terms in DALoss. Tab. 5 shows the efficacy of the different components in DALoss, which suggests that: (1) DALoss yields an improvement of 1.8 points to the model. Specifically, the mAP increases from 45.5 to 48.3 on AP_L, indicating a significant improvement for large objects. (2) The aligning weights for classification loss alone can also provide a marginal performance gain for the model. (3) Simply adding a mask IoU branch to the model does not enhance the overall performance. However, when incorporating the aligning weights for the segmentation loss, the mAP is increased. (4) Interestingly, the mask IoU branch can help improve the AP for objects of different scales. This is due to the mask IoU branch helping to select good pseudo-labels for training.

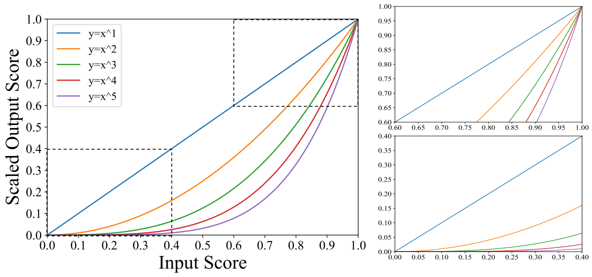

Figure 3. **Visualization of the loss alignment with different hyper-parameters.** When changing the hyper-parameter from 1 to 5, the input classification score or mask IoU will be more properly adjusted by their quality, *i.e.*, the difference between high-&low-quality mask or class will be enlarged.

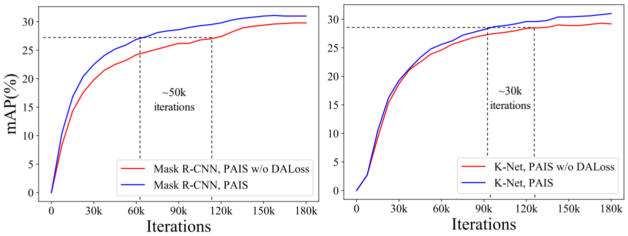

Figure 4. **Model convergence speed.** DALoss leads to faster model convergence for training.

τ_{cls}	AP	$\text{AP}_{0.5}$	$\text{AP}_{0.75}$	AP_S	AP_M	AP_L
0.35	31.1	50.6	32.4	12.3	33.3	48.3
0.50	30.6	50.0	31.9	11.2	32.6	47.8
0.65	29.9	49.1	31.3	10.1	32.1	47.4

Table 7. **Results of different classification score threshold τ_{cls}.** The mask IoU threshold τ_{iou} is set to 0.3.

τ_{iou}	AP	$\text{AP}_{0.5}$	$\text{AP}_{0.75}$	AP_S	AP_M	AP_L
0.3	31.1	50.6	32.4	12.3	33.3	48.3
0.5	30.8	50.3	31.9	12.0	32.9	47.9
0.7	30.6	50.0	31.2	11.6	32.3	47.1

Table 8. **Results of different mask IoU threshold τ_{iou}.** The classification score threshold τ_{cls} is set to 0.35.

Ratio	AP	$\text{AP}_{0.5}$	$\text{AP}_{0.75}$	AP_S	AP_M	AP_L
1:1	29.0	47.7	31.0	11.0	31.6	45.5
1:2	30.1	48.8	32.2	11.5	32.4	47.1
1:3	31.1	50.6	32.4	12.3	33.3	48.3
1:4	31.3	50.9	32.6	12.4	33.5	48.7

Table 9. **Different ratio of labeled images and unlabeled images in a batch.** The performance saturates when the ratio of labeled and unlabeled images is set to 1:4.

Hyper-parameters in DALoss. We investigated the effect of hyper-parameters, α and β, used in DALoss. To give equal weight to both classification and segmentation, we set $\alpha = \beta$. The results are shown in Tab. 6, indicating that larger values of α and β lead to better performance. We

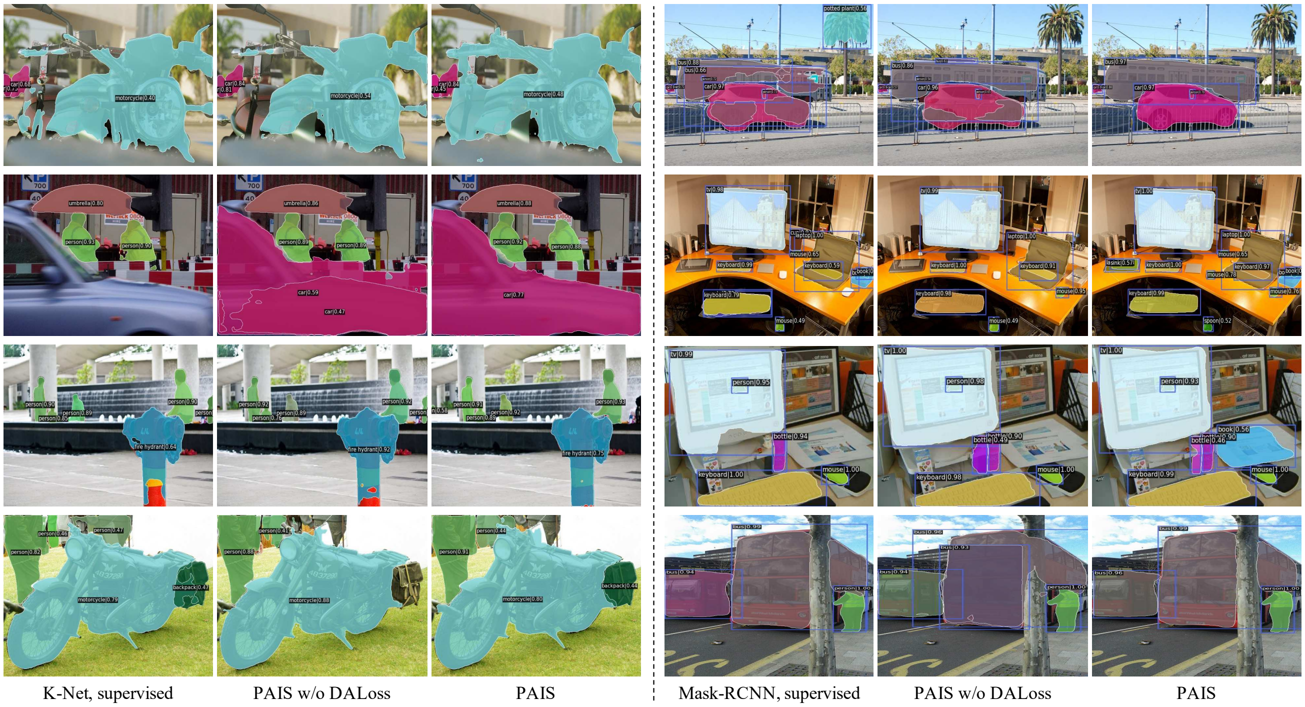

Figure 5. **Visualization of the predictions from different models.** From the examples, we can clearly show the benefits introduced by pixel-level information from PAIS and DALoss.

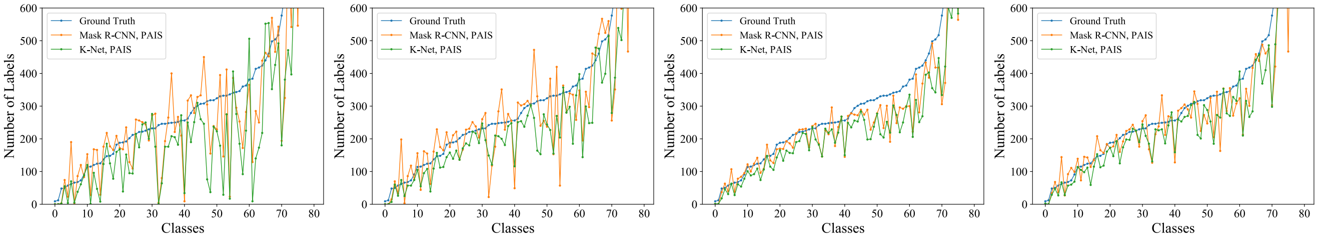

Figure 6. **Influence of imbalanced classes to semi-supervised training**. The models are trained under 10% labeled images for the results. We randomly sampled 10k images from the rest of COCO `train2017` to draw these figures. As the training progresses, the predictions from the teacher model are gradually fitted to the distribution of the number of labels.

also analyzed the functions of different hyper-parameters used to adjust the input score, as shown in Fig. 3. The figures illustrate that increasing the hyper-parameter enlarges the difference between high- and low-quality masks or classes, adjusting the input score more appropriately based on their quality. However, when the hyper-parameter is set too large, the low-quality masks or classes are significantly constrained, potentially affecting the generalization of the model as noise is removed entirely during training.

Model convergence speed. In Fig. 4, we analyze the convergence speed of the models. The results demonstrate that DALoss can expedite the model training in terms of convergence rate. For instance, Mask-RCNN (PAIS) achieves a convergence speed that is approximately 2 times faster than that of Mask-RCNN (PAIS w/o DALoss). Moreover, by comparing the results between Mask-RCNN (PAIS) and K-Net (PAIS), we observe that DALoss may be particularly effective for box-dependent instance segmentation frameworks, which facilitates rapid convergence.

Different thresholds for classification scores and mask IoUs. We observed from Tab. 7 and Tab. 8 that increasing the thresholds for classification score and mask IoU leads to a decrease in performance. This suggests that DALoss is able to make use of misaligned pseudo-labels instead of simply filtering them out. By allowing the model to learn from these potentially noisy labels, it is able to better handle situations where the alignment between labeled and unlabeled data is not perfect. This may result in improved generalization to new, unseen data, as the model has learned to adapt to the presence of noisy labels.

Different ratios of labeled and unlabeled images per batch. Tab. 9 shows that increasing the ratio of unlabeled

τ_{cls}, τ_{iou}	0.35, 0.3	0.50, 0.3	0.65, 0.3	0.35, 0.5	0.35, 0.7
w/o DALoss	29.3	29.0	25.3	29.3	29.0
w/ DALoss	31.1	30.6	29.9	30.8	30.6

Table 10. DALoss w.r.t. thresholds tuning. DALoss consistently outperforms threshold tuning in all settings.

Method	1%	5%	10%	100%
K-Net, supervise	11.6	22.3	26.5	38.4
EMA w/o DALoss	17.8*(+6.2)*	25.4*(+3.1)*	29.3*(+2.8)*	39.4*(+1.0)*
EMA w/ DALoss	19.8*(+8.2)*	27.5*(+5.2)*	31.0*(+4.5)*	40.8*(+2.4)*
Performance gain	*+2.0*	*+2.1*	*+1.7*	*+1.4*
M-RCNN, supervise	11.5	22.4	27.1	37.5
EMA w/o DALoss	20.1*(+8.6)*	27.4*(+5.0)*	29.8*(+2.7)*	38.4*(+0.9)*
EMA w/ DALoss	21.1*(+9.6)*	29.3*(+6.9)*	31.0*(+3.9)*	39.5*(+2.0)*
Performance gain	*+1.0*	*+1.9*	*+1.2*	*+1.1*

Table 11. Comparison of EMA with DALoss. The performance improvements are not mainly attributed to EMA.

images in a batch can improve the model's performance. The results indicate that performance saturates when the ratio is set to 1:4, suggesting that adding too many unlabeled images may lead to diminishing returns.

Visualizations. We visualize the outputs of different models, including K-Net (supervised), K-Net (PAIS w/o DALoss), K-Net (PAIS), Mask-RCNN (supervised), Mask-RCNN (PAIS w/o DALoss), and Mask-RCNN (PAIS), in Fig. 5. The results show that: (1) PAIS can improve the recall for most instances in the supervised model, but the quality of the predictions is not guaranteed. (2) The use of DALoss in PAIS helps to improve the quality of predictions in terms of mask and classification.

Influence of imbalanced classes. In Fig. 6, we investigate the effect of imbalanced classes on PAIS. We plot the number of labels obtained from the ground truth and the teacher models for Mask-RCNN and K-Net with PAIS at different iterations, namely 32k, 64k, 120k, and 180k. The results indicate that the predicted pseudo-labels gradually conform to the distribution of the imbalanced labels.

5. Discussion

DALoss w.r.t. Thresholds Tuning. As shown in Tab. 10, we conduct an ablation study by removing DALoss from our method in Tabs. 7 and 8. The results show that DALoss consistently outperforms threshold tuning in all settings, which validates that DALoss is indeed more effective than tuning the thresholds.

Comparison of EMA with DALoss. We show that the performance improvements are not mainly attributed to EMA. First, we summarize the results of Tabs. 3 and 4 in Tab. 11. We can find that DALoss achieves a larger performance gain than EMA in 100% COCO (+1.4, +1.1 *vs.* +1.0, +0.9). Second, EMA alone cannot selectively utilize noisy pseudo-labels for better learning. DALoss solves this problem and leads to further improvement over EMA.

Discussion on Score Filtering. We have carefully chosen the values of τ_{cls}, τ_{iou}, based on Fig. 1(c)(d), which illustrates that decreasing the thresholds will generate more noisy but informative pseudo-labels. Therefore, we use low thresholds to obtain such pseudo-labels, which is different from previous methods that need to adjust thresholds to filter them out. This simplifies the tuning of the thresholds. We believe that the experiments in Tabs. 7, 8, and 10 are sufficient to show that lower thresholds are suitable for DALoss.

Generality to Other Segmentation Tasks. DALoss can be applied to other semi-supervised segmentation tasks, as the noisy pseudo-labels are common in semi-supervised setting. To show the generality, we apply DALoss to panoptic segmentation, which is a more challenging task that requires both instance and semantic segmentation. We report a preliminary result under 10% COCO. The initial result shows that DALoss can improve the PQ performance from 36.8% PAIS w/o DALoss to 37.3% PAIS w/ DALoss. We remain more experiments in our future work.

Discussion on Large Segment Everything Models. We envision that the future of image segmentation will not only aim to segment everything, but also to provide fine-grained text descriptions for the segmented regions. However, the recently proposed models such as SAM [19] and SEEM [55] either lack labeled semantic information or demand large amounts of labeled semantic data for training. Therefore, we believe that semi-supervised learning will be a crucial solution to leverage abundant unlabeled data and reduce the labeling burden.

6. Conclusion

In this paper, we presented a novel PAIS framework for semi-supervised instance segmentation. To address the misalignment between classification score and mask quality, we introduced a dynamic aligning loss (DALoss), which aligns the classification loss term and the segmentation loss term based on the quality of different pseudo-labels. Our experimental results demonstrate the effectiveness of the proposed PAIS framework. Specifically, when the amount of labeled data is extremely limited, our pipeline equipped with PAIS and DALoss achieves superior performance for instance segmentation. We believe that PAIS can serve as a strong baseline for future research on semi-supervised instance segmentation. We hope our work can inspire further exploration in this exciting research direction.

Acknowledgements. This work was supported by National Key R&D Program of China (2022ZD0118202), National Science Fund for Distinguished Young Scholars (No.62025603), National Natural Science Foundation of China (No.U21B2037, No.U22B2051, No.62176222, No.62176223, No.62176226, No.62072386, No.62072387, No.62072389, No.62002305 and No.62272401), and Natural Science Foundation of Fujian Province of China (No.2021J01002, No.2022J06001).

References

[1] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. *Advances in Neural Information Processing Systems*, 2014. 2

[2] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. In *International Conference on Learning Representations*, 2020. 2

[3] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 2

[4] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[5] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 2021. 2, 3

[6] 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/CVF Conference on Computer Vision and Pattern Recognition*, 2016. 5

[7] Zihang Dai, Zhilin Yang, Fan Yang, William W Cohen, and Russ R Salakhutdinov. Good semi-supervised learning that requires a bad gan. *Advances in Neural Information Processing Systems*, 2017. 2

[8] Bert De Brabandere, Davy Neven, and Luc Van Gool. Semantic instance segmentation with a discriminative loss function. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, 2017. 2

[9] Dominik Filipiak, Andrzej Zapała, Piotr Tempczyk, Anna Fensel, and Marek Cygan. Polite teacher: Semi-supervised instance segmentation with mutual learning and pseudo-label thresholding. *arXiv preprint arXiv:2211.03850*, 2022. 3

[10] Geoffrey French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. In *International Conference on Learning Representations*, 2018. 2

[11] Naiyu Gao, Yanhu Shan, Yupei Wang, Xin Zhao, Yinan Yu, Ming Yang, and Kaiqi Huang. Ssap: Single-shot instance segmentation with affinity pyramid. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 2

[12] Ruohao Guo, Dantong Niu, Liao Qu, and Zhenbo Li. Sotr: Segmenting objects with transformers. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 3

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 1, 2, 3, 5, 6

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016. 5

[15] Jie Hu, Liujuan Cao, Yao Lu, ShengChuan Zhang, Ke Li, Feiyue Huang, Ling Shao, and Rongrong Ji. Istr: End-to-end instance segmentation via transformers. *arXiv preprint arXiv:2105.00637*, 2021. 2

[16] Jie Hu, Linyan Huang, Tianhe Ren, Shengchuan Zhang, Rongrong Ji, and Liujuan Cao. You only segment once: Towards real-time panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17819–17829, 2023. 2

[17] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 3, 4

[18] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. *Advances in Neural Information Processing Systems*, 2019. 5

[19] 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. *arXiv preprint arXiv:2304.02643*, 2023. 9

[20] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations*, 2017. 2

[21] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *International Conference on Machine Learning (Workshop)*, 2013. 2

[22] Youngwan Lee and Jongyoul Park. Centermask: Real-time anchor-free instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2

[23] Xinyang Li, Jie Hu, Shengchuan Zhang, Xiaopeng Hong, Qixiang Ye, Chenglin Wu, and Rongrong Ji. Attribute guided unpaired image-to-image translation with semi-supervised learning. *arXiv preprint arXiv:1904.12428*, 2019. 2

[24] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 2

[25] 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 *European Conference on Computer Vision*, 2014. 5

[26] Shu Liu, Jiaya Jia, Sanja Fidler, and Raquel Urtasun. Sgn: Sequential grouping networks for instance segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 2

[27] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2

[28] Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. *International Conference on Learning Representations*, 2021. 2, 3, 5

[29] Yao Lu, Zhiyi Chen, Zehui Chen, Jie Hu, Liujuan Cao, and Shengchuan Zhang. Candy: Category-kernelized dynamic convolution for instance segmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023. 2

[30] Wenfeng Luo and Meng Yang. Semi-supervised semantic segmentation via strong-weak dual-branch network. *Springer International Publishing eBooks*, 2020. 5

[31] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 638–647, 2022. 2

[32] Yiwei Ma, Xiaioqing Zhang, Xiaoshuai Sun, Jiayi Ji, Haowei Wang, Guannan Jiang, Weilin Zhuang, and Rongrong Ji. X-mesh: Towards fast and accurate text-driven 3d stylization via dynamic textual guidance. *arXiv preprint arXiv:2303.15764*, 2023. 2

[33] Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 2

[34] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. *arXiv: Computer Vision and Pattern Recognition*, 2020. 5

[35] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 1, 5

[36] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. *Advances in Neural Information Processing Systems*, 2015. 2

[37] Tianhe Ren, Shilong Liu, Feng Li, Hao Zhang, Ailing Zeng, Jie Yang, Xingyu Liao, Ding Jia, Hongyang Li, He Cao, Jianan Wang, Zhaoyang Zeng, Xianbiao Qi, Yuhui Yuan, Jianwei Yang, and Lei Zhang. detrex: Benchmarking detection transformers, 2023. 2

[38] Tianhe Ren, Jianwei Yang, Shilong Liu, Ailing Zeng, Feng Li, Hao Zhang, Hongyang Li, Zhaoyang Zeng, and Lei Zhang. A strong and reproducible object detector with only public datasets, 2023. 2

[39] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in Neural Information Processing Systems*, 2016. 2

[40] Henry Scudder. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 1965. 2

[41] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 2020. 2

[42] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020. 2, 5

[43] Yihe Tang, Weifeng Chen, Yijun Luo, and Yuting Zhang. Humble teachers teach better students for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2

[44] 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*, 2017. 2, 3, 5

[45] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. Solo: Segmenting objects by locations. In *European Conference on Computer Vision*, 2020. 2, 3

[46] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. *Advances in Neural Information Processing Systems*, 2020. 2, 3

[47] Zhenyu Wang, Yali Li, and Shengjin Wang. Noisy boundaries: Lemon or lemonade for semi-supervised instance segmentation? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 3, 5, 6

[48] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 2020. 2

[49] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2

[50] 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 International Conference on Computer Vision*, 2021. 1, 2, 3

[51] Qize Yang, Xihan Wei, Biao Wang, Xian-Sheng Hua, and Lei Zhang. Interactive self-training with mean teachers for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2

[52] Jialin Yuan, Chao Chen, and Li Fuxin. Deep variational instance segmentation. In *Advances in Neural Information Processing Systems*, 2020. 2

[53] Wenwei Zhang, Jiangmiao Pang, Kai Chen, and Chen Change Loy. K-net: Towards unified image segmentation. *Advances in Neural Information Processing Systems*, 2021. 1, 2, 3, 5, 6

[54] Qiang Zhou, Chaohui Yu, Zhibin Wang, Qi Qian, and Hao Li. Instant-teaching: An end-to-end semi-supervised object detection framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2

[55] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, YongJae Lee, Madison Madison, Microsoft Research, Redmond Hkust, Microsoft Cloud, and Ai Ai. Segment everything everywhere all at once. 9