Pseudo-label Alignment for Semi-supervised Instance Segmentation

Jie Hu\textsuperscript{1}, Chen Chen\textsuperscript{1}, Liujuan Cao\textsuperscript{1}, Shengchuan Zhang\textsuperscript{1}, Annan Shu\textsuperscript{2}, Guannan Jiang\textsuperscript{2}, and Rongrong Ji\textsuperscript{1}

\textsuperscript{1}Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University
\textsuperscript{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: \url{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 \cite{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 Cityscapes 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 \cite{13} and K-Net \cite{53}, respectively. This outperforms the current state-of-the-art model, NoisyBoundary \cite{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,
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
In this paper, we investigate PAIS in both box-free and 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

1: Initialize teacher and student models randomly.
2: Repeat
3: Apply weak augmentation to unlabeled images.
4: Obtain pseudo-labels \( \{m_i, b_i, c_i, s_i\}_{i=1, \ldots, N} \) from teacher model via thresholds \( \tau_{cls}, \tau_{iou} \).
5: Apply strong augmentation to unlabeled images.
6: Obtain predictions \( \bar{m}, \bar{b}, \bar{c} \) for unlabeled images.
7: Calculate \( L_t \) for unlabeled images by Eq. 5.
8: Inference labeled images by student model.
9: Calculate \( L_l \) for labeled images.
10: Train student model with the loss \( L_{semi} \) in Eq. 1.
11: Update teacher model via EMA.
12: 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 \( \tau_{cls} \) and a mask IoU threshold \( \tau_{iou} \). The resulting set of pseudo-labels includes \( N \) elements, each containing a mask \( m_i \in \mathbb{R}^{H \times W} \), a bounding box \( b_i \in \mathbb{R}^4 \), a classification score \( c_i \in \mathbb{R}^L \) (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:

\[
L_{semi} = \lambda_1 L_t + \lambda_u L_u,
\]

where \( \lambda_1 \) and \( \lambda_u \) 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:

\[
L_u = \frac{1}{N_{pos}} \sum_{i=1}^{N_{pos}} \left( \lambda_b L_b(b_{\nu(i)}, b_{\mu(i)}) + \lambda_c L_c(\bar{c}_{\nu(i)}, c_{\mu(i)}) + \lambda_m L_m(\bar{m}_{\nu(i)}, m_{\mu(i)}) \right)
\]

where \( L_b \) includes the box IoU loss and the L1 loss, \( L_c \) denotes the cross-entropy loss, \( L_m \) is the dice loss. In Eq. 2, \( \bar{c}_i, \bar{b}_i, \bar{m}_i \) represent predictions from the student model.

The pseudo-scores \( c_{\mu(i)} \) for classification are converted to one-hot vectors \( c_{\mu(i)}^h \) for the category with the highest score.

\[
\nu(i) = \arg \min_{\nu \in \{1, \ldots, N\}} \left( L_b(b_i, b_{\nu}) \right).
\]

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

\[
\nu^* = \arg \min_{\nu} \sum_{i=1}^{t} \left( \lambda_b L_b(b_{\nu(i)}, b_i) + \lambda_c L_c(\bar{c}_{\nu(i)}, c_i^h) + \lambda_m L_m(\bar{m}_{\nu(i)}, m_i^b) - s_i \right).
\]

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:

\[
L_{w}^{DA} = \frac{1}{N_{pos}} \sum_{i=1}^{N_{pos}} \left( \lambda_b L_b(b_{\nu(i)}, b_{\mu(i)}) + \lambda_c L_c(\bar{c}_{\nu(i)}, c_{\mu(i)}^h) + \lambda_m L_m(\bar{m}_{\nu(i)}, m_{\mu(i)}^b) \right)
\]

where \( c_{\mu(i)}^h \) denotes the highest classification score, and \( s_{\mu(i)} \) denotes the mask IoU from the \( \mu(i) \)-th pseudo-label,
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 $\lambda_l$ and $\lambda_u$, 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 $\tau_{cls}$ and $\tau_{iou}$ 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 $\lambda_b=2.0$, $\lambda_c=4.0$, and $\lambda_m=1.0$, 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-

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

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 NoisyBoundaries [47].

\[ \alpha, \beta \text{ are the hyper-parameters. Specifically, Eq. 5 adjusts the weights for the pseudo-labels conditioned on their qualities, i.e., dynamically dependent on the input images. For instance, for a pseudo-label with a low classification score and high mask IoU, DALoss encourages the segmentation loss while constraining the classification loss.} \]

4.4. 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.

4.5. 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-
To demonstrate the effectiveness of PAIS, we constructed pseudo-labels, which matches our goal to utilize noisy labeled images. The COCO dataset can provide more diverse and informative pseudo-labels, while the Cityspaces dataset only has 8 instance categories. This implies that the COCO dataset has 80 instance categories, while the Cityspaces dataset has 8 instance categories. This allows us to solve the noisy pseudo-label problem. The COCO dataset validates the effectiveness of our method in the fully-supervised setting.

We randomly sampled 1%, 5%, and 10% of the labeled images from the COCO 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 NoisyBoundaries [47]. We can see that PAIS achieves good performance on both box-free and box-independent instance segmentation frameworks.

<table>
<thead>
<tr>
<th>Method</th>
<th>1% COCO</th>
<th>5% COCO</th>
<th>10% COCO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AP</td>
<td>AP₀.₅</td>
<td>AP₀.₇₅</td>
</tr>
<tr>
<td><strong>Box-free Instance Segmentation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K-Net [53], supervised</td>
<td>8.03±0.25 16.33±0.31 7.00±0.25</td>
<td>17.4±0.22 30.08±0.28 16.83±0.21</td>
<td>21.63±0.21 37.43±0.38 21.8±0.20</td>
</tr>
<tr>
<td>K-Net¹ [53], supervised</td>
<td>11.63±0.05 22.30±0.08 10.95±0.10</td>
<td>22.28±0.05 38.54±0.09 22.7±0.07</td>
<td>26.53±0.12 44.87±0.15 27.2±0.17</td>
</tr>
<tr>
<td>K-Net, PAIS w/o DALoss</td>
<td>17.77±0.06 32.17±0.15 17.53±0.12</td>
<td>25.40±0.08 43.12±0.05 26.05±0.13</td>
<td>29.30±0.07 48.64±0.05 30.53±0.13</td>
</tr>
<tr>
<td>K-Net, PAIS</td>
<td>19.78±0.10 35.48±0.15 19.65±0.06</td>
<td>27.53±0.06 46.53±0.06 28.70±0.10</td>
<td>31.04±0.06 50.50±0.07 32.34±0.05</td>
</tr>
</tbody>
</table>

| **Box-dependent Instance Segmentation** |
| Mask-RCNN [13], supervised* | 3.5 | -  | -    | 17.3 | -  | -    | 22.0 | -  | -    |
| Mask-RCNN¹ [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 NoisyBoundaries [47]. We can see that PAIS achieves good performance on both box-free and box-independent instance segmentation frameworks.

<table>
<thead>
<tr>
<th>Method</th>
<th>AP</th>
<th>AP₀.₅</th>
<th>AP₀.₇₅</th>
<th>AP</th>
<th>AP₀.₅</th>
<th>AP₀.₇₅</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Box-free Instance Segmentation</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K-Net [53], supervised</td>
<td>37.8</td>
<td>60.3</td>
<td>39.9</td>
<td>16.9</td>
<td>41.2</td>
<td>57.5</td>
</tr>
<tr>
<td>K-Net¹ [53], supervised</td>
<td>38.4</td>
<td>61.4</td>
<td>40.3</td>
<td>17.6</td>
<td>41.8</td>
<td>58.0</td>
</tr>
<tr>
<td>K-Net, PAIS w/o DALoss</td>
<td>39.4</td>
<td>62.2</td>
<td>41.6</td>
<td>18.5</td>
<td>42.8</td>
<td>59.2</td>
</tr>
<tr>
<td>K-Net, PAIS</td>
<td>40.8</td>
<td>63.5</td>
<td>43.3</td>
<td>19.2</td>
<td>44.4</td>
<td>61.4</td>
</tr>
</tbody>
</table>

| **Box-dependent Instance Segmentation** |
| Mask-RCNN [13], supervised* | 37.1 | 58.5 | 39.7  | 18.7 | 39.6 | 53.9  |
| Mask-RCNN¹ [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 unlabeled2017 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 Cityscapes dataset only has 8 instance categories. This implies that the COCO dataset can provide more diverse and informative abusive 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-Net¹ (supervised) and Mask-RCNN¹ (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-Net¹ (supervised) and Mask-RCNN¹ (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 to 19.78 and 21.12, respectively, resulting in approximately +8.15 and +9.58 points improvement to the performance. Additionally, the improvement over AP₀.₅ and AP₀.₇₅ 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.
Effectiveness of different loss terms in DALoss. Tab. 5 shows 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.

<table>
<thead>
<tr>
<th>α, β</th>
<th>AP</th>
<th>AP₀.5</th>
<th>AP₀.75</th>
<th>AP₅₀</th>
<th>AP₇₅</th>
<th>AP₉₅</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>29.9</td>
<td>49.1</td>
<td>31.0</td>
<td>10.9</td>
<td>32.1</td>
<td>47.0</td>
</tr>
<tr>
<td>2</td>
<td>30.2</td>
<td>49.5</td>
<td>31.1</td>
<td>11.1</td>
<td>32.2</td>
<td>47.1</td>
</tr>
<tr>
<td>3</td>
<td>30.5</td>
<td>50.1</td>
<td>31.7</td>
<td>11.6</td>
<td>32.9</td>
<td>47.6</td>
</tr>
<tr>
<td>4</td>
<td>31.1</td>
<td>50.6</td>
<td>32.4</td>
<td>12.3</td>
<td>33.3</td>
<td>48.3</td>
</tr>
</tbody>
</table>

Table 6. Hyper-parameters in DALoss. Considering that the classification and segmentation are both important, we simply set α = β to investigate the influence of hyper-parameters.

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 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.

Table 7. Results of different classification score threshold τcls. The mask IoU threshold τiou is set to 0.3.

Table 8. Results of different mask IoU threshold τiou. The classification score threshold τcls is set to 0.35.

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.
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
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), 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 $\tau_{cls}$, $\tau_{iou}$, 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), Natural Natural Science Foundation of China (No.2121B2037, No.6212B2015, No.62176222, No.62176223, No.62176226, No.62072386, No.62072387, No.62072389, No.620022305 and No.62272401), and Natural Science Foundation of Fujian Province of China (No.2021J01002, No.2022J06001).

<table>
<thead>
<tr>
<th>$\tau_{cls}$, $\tau_{iou}$</th>
<th>0.35, 0.3</th>
<th>0.50, 0.3</th>
<th>0.65, 0.3</th>
<th>0.35, 0.5</th>
<th>0.35, 0.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>w/o DALoss</td>
<td>29.3</td>
<td>29.0</td>
<td>25.3</td>
<td>29.3</td>
<td>29.0</td>
</tr>
<tr>
<td>w/ DALoss</td>
<td>31.1</td>
<td>30.6</td>
<td>29.9</td>
<td>30.8</td>
<td>30.6</td>
</tr>
</tbody>
</table>

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

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

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


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


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

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


[51] Wenwei Zhang, Jiawei Li, Chao Cao, and Binghui Li. Non-Local Transformer: A Simple and Effective Model for Image Segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021. 2


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