S4OD: Semi-Supervised learning for Single-Stage Object Detection

Yueming Zhang¹,², Xingxu Yao², Chao Liu², Feng Chen², Xiaolin Song², Tengfei Xing², Runbo Hu², Hua Chai², Pengfei Xu², and Guoshan Zhang¹

¹ Tianjin University
{seife, zhanggs}@tju.edu.cn
² Didi Chuxing

Abstract. Single-stage detectors suffer from extreme foreground-background class imbalance, while two-stage detectors do not. Therefore, in semi-supervised object detection, two-stage detectors can deliver remarkable performance by only selecting high-quality pseudo labels based on classification scores. However, directly applying this strategy to single-stage detectors would aggravate the class imbalance with fewer positive samples. Thus, single-stage detectors have to consider both quality and quantity of pseudo labels simultaneously. In this paper, we design a dynamic self-adaptive threshold (DSAT) strategy in classification branch, which can automatically select pseudo labels to achieve an optimal trade-off between quality and quantity. Besides, to assess the regression quality of pseudo labels in single-stage detectors, we propose a module to compute the regression uncertainty of boxes based on Non-Maximum Suppression. By leveraging only 10% labeled data from COCO, our method achieves 35.0% AP on anchor-free detector (FCOS) and 32.9% on anchor-based detector (RetinaNet).

Keywords: Single Stage, Object Detection, Semi-supervised Learning, Sampling Imbalance

1 Introduction

The performance of deep learning networks benefit from tremendous data [7]. However, it’s expensive and intractable to obtain a large-scale human-labeled dataset for a new application scenario. This issue has promoted semi-supervised learning to become the research topic that receives substantial interests in various fields [6,13,37,35]. For object detection tasks with significant application value yet high annotation cost, it is of great significance to research on semi-supervised object detection.

Most previous studies focus on the two-stage detectors [11,19,26,39,30,35]. They state that their methods could be applied to single-stage object detectors in theory, but few empirical results can systematically support this. As proved in [10], the major difference between single- and two-stage detectors is the class balance difference. Specifically, in two-stage detectors, region proposal network
Fig. 1. (a): Compared with Faster-RCNN, FCOS suffers from a serious sampling imbalance (orange triangle) for pseudo labels which further leads to poor performance (blue triangle). Orange lines are the ratio of positives and negatives in FCOS and that in Faster-RCNN on pseudo labels, and the blue lines are the mAP curves on COCO-val correspondingly. (b): The distribution of classification score on COCO-val, only predictions with IoU $> 0.9$ are counted.

(RPN) filters out most of background samples, and a fixed ratio is adopted in the next stage to maintain a sampling balance between foregrounds and backgrounds. However, single-stage detectors directly make dense predictions covering all space locations in each scale. Only a small fraction of predictions are assigned as positive samples. This leads to a severe imbalance of positive and negative samples. In semi-supervised learning, the improper pseudo labels selection strategy would aggravate this imbalance. For example, as shown in Fig 1(a), directly using the same setting from two-stage detectors in single-stage detectors would drastically sharpen the number of positive samples, pushing the ratio of positives and negatives towards zero. Such imbalance forces the detectors to make predictions with an inductive bias on the background class, and therefore limits the performance of semi-supervised single-stage detectors.

To alleviate the sampling imbalance of single-stage detectors in semi-supervised learning, we focus on the strategy of selecting pseudo labels. Conventional methods [35, 19] use an ablation study to determine the strategy of pseudo label filtering, which is complicated and unexplainable. Most importantly, this strategy lowers the practical value of these semi-supervised learning methods. For example, [35] and [19] both use Faster-RCNN [24], while the most proper thresholds obtained from the ablation study are not identical (0.9 and 0.7). This case motivates us to find a more universal pseudo label selecting strategy. To this end, we first compare the classification distribution of final predictions between the single- and two-stage detectors. As shown in Fig 1(b), the classification score of single-stage detectors is significantly lower than that of two-stage detectors. This gap indicates that transferring the selective strategy of two-stage detectors to the single-stage detectors would reduce the number of pseudo labels and further
lead to a more severe sampling imbalance, although the quality of pseudo labels is still guaranteed. In contrast, choosing a loose strategy would increase the number of pseudo labels and alleviate this imbalance, but more noises are brought and the quality of pseudo labels cannot be promised. Therefore, inspired by this insight, we propose a dynamic self-adaptive threshold (DSAT) which can automatically determine the selective strategy considering both quality and quantity of pseudo labels. More importantly, this strategy can dynamically select pseudo labels which are aligned with the model performance during the training process.

As another part of object detection, the regression quality of pseudo labels is of equal importance as the classification score. However, it is hard to evaluate the regression quality in continuous space. To avoid the impact caused by the low-quality pseudo labels, Liu et al. [19] abandon the regression loss when computing the unsupervised loss. Xu et al. [35] find that the regression results are related to the uncertainty which is computed by the BoxJitter module. However, the BoxJitter module relies on the Region-CNN (RCNN) to process the jittered proposals, which means that it is infeasible for single-stage detectors. To address this issue, we define the standard deviation of the redundant bounding boxes in non-maximum suppression (NMS) as the corresponding box uncertainty and choose the pseudo labels with lower uncertainty as the regression target. The proposed module is named NMS-UNC as it is integrated with the original NMS module.

In summary, we propose a semi-supervised learning framework for single-stage detectors (S4OD). Our contributions can be summarized as follows:

- We reveal the importance of pseudo label selecting strategy in semi-supervised object detection and propose a DSAT strategy to dynamically determine the strategy which takes both quality and quantity of pseudo labels into consideration.
- To assess the quality of box regression in single-stage detectors, we develop a NMS-UNC module which computes the uncertainty of boxes based on the original NMS.
- Experiments of using different percentages of labeled data on MS-COCO dataset show our method achieves a significant improvement (e.g. +10.1% AP using 1% labeled data) than fully-supervised learning.

2 Related Works

2.1 Object Detection

In general, CNN-based object detectors can be split into two categories: proposal-based two-stage detectors [24,15,9,15] and single-stage detectors [30,8,23,18,29,16]. Two-stage detectors rely on the sparse sampling of RPN to filter out most background samples and further refine the class balanced proposals. However, single-stage detectors use a single convolution network to directly make dense predictions, inevitably leading to the sampling imbalance. Single-stage detectors
are more efficient while the two-stage detectors take the lead in terms of accuracy. Another paradigm is transformer-based method \cite{21,20,40}, these detectors achieve excellent performance relying on the strong feature extracting capacity of transformer. In this paper, we focus on two typical single-stage detectors: FCOS \cite{29} for anchor-free method and RetinaNet \cite{16} for anchor-based method.

\section{2.2 Semi-Supervised Learning in Object Detection}

Applying semi-supervised learning in object detection is more challenging than that in classification \cite{22,25,34,33,2}, because object detection combines both classification and regression. CSD \cite{11} explicitly applies the consistency regularization on both labeled and unlabeled data for semi-supervised object detection. STAC \cite{26} is a semi-supervised learning framework based on self-training strategy and augmentation-driven consistency regularization. Instant-Teaching \cite{39} is designed to be a two-model co-rectify framework for increasing the accuracy of pseudo labels during the training process. Wang et al. \cite{32} divide unlabeled data according to the difficulty level and utilizes data from different levels in different training phases. ISD \cite{12} proposes a novel interpolation-based semi-supervised learning algorithm for object detection and conducts experiments on SSD \cite{18} network, but only anchor-based single-stage detector is considered. In this paper, we employ the teacher-student mutual learning framework in an end-to-end manner without redundant implementation, and conduct experiments on both anchor-free and anchor-based single-stage object detectors. Our model significantly improves the application potential of semi-supervised learning in single-stage object detector.

\section{3 Preliminary of Semi-Supervised Learning}

\subsection{3.1 Mutual Learning Framework}

Mainstream semi-supervised methods \cite{19,27,35} for object detection use the teacher-student mutual learning framework, consisting of a student model $\theta^S$ and a teacher model $\theta^T$. For labeled data $D^l$, the student model $\theta^S$ uses its ground truth to compute loss $L_{sup}$ for gradient descent:

$$L_{sup} = \frac{1}{N_{pos}} \sum_{x,y} L_{cls}(p_{x,y}, c_{x,y}) + \frac{1}{N_{pos}} \sum_{i} L_{reg}(t_i, b_i)$$

where $N_{pos}$ is the number of positive samples, $p_{x,y}$ and $c_{x,y}$ refer to the predicted logit and ground-truth at location $(x, y)$. $t_i$ denotes the predicted regression result of the $i$-th positive sample, while $b_i$ is the corresponding target.

As for unlabeled data, the teacher model $\theta^T$ firstly inferses on the weak-augmented unlabeled data $D_{weak}^u$. Then a classification score threshold $\sigma_{cls}$ is determined to filter out low-quality results. The reserved high-quality prediction is regarded as the ground truth (pseudo labels) for the strong-augmented
unlabeled data $D_{\text{strong}}^u$. Afterwards, the student model computes unsupervised loss $L_{\text{unsup}}$ as Eq. (2) for unlabeled data and updates parameters $\theta^S$ through gradient descent. Finally, the teacher model $\theta^T$ is updated by the student model $\theta^S$ with exponential moving average (EMA) [28].

$$L_{\text{unsup}} = \frac{1}{N_{\text{pos}}} \sum_{x,y} L_{\text{cls}}(p_{x,y}, c^*_{x,y}) + \frac{1}{N_{\text{pos}}} \sum_{i} L_{\text{reg}}(t_i, b^*_i)$$ (2)

where $c^*$ and $b^*$ are the categories and boxes of positive samples. These positive samples are assigned from the reserved pseudo labels with higher classification scores than threshold $\sigma_{\text{cls}}$.

### 3.2 Difference between Single- and Two-Stage Detectors

Two-stage detectors include a RPN to generate proposals and a cascaded network to predict classification scores and refine regression results of each proposal, while single-stage object detectors directly generate dense outputs to predict the classification scores and regression results for each location.

In order to have a better insight into the semi-supervised implementation difference between single- and two-stage detectors, we visualise the classification confidence distribution of single- and two-stage detectors in Fig.1(b). As shown in Fig.1(b), although the performances are comparable (37.4% and 37.9% mAP), there is a significant difference in classification score distribution between the single- and two-stage detectors. The classification confidence in the two-stage detectors (Faster-RCNN [24]) is generally higher than that in the single-stage (FCOS [29]). Specifically, most classification scores of the proposals in Faster-RCNN locate in [0, 1.0), while that in FCOS locate in the range of [0.6, 0.8]. For two-stage detectors, with the filtering and sampling capacity of RPN, the second stage could generate strong confidence scores on plenty of reserved high-quality proposals. Hence, two-stage detectors could benefit from choosing a more strict selecting strategy of pseudo labels. This strategy ensures the quality of pseudo labels without losing too much pseudo labels. However, single-stage detectors make predictions on the imbalanced samples, so the confidence scores are not as high as that in two-stage. This difference indicates that the selective strategy of pseudo labels in single-stage detectors must consider both the quality and quantity of them. A more strict selective strategy could ensure the quality of pseudo labels but the amount of them decreased greatly. This would exacerbate the inherent imbalance of positive and negative samples. The imbalanced sampling ratio leads negative samples dominating gradient and push the network to generate predictions with bias on background. As the teacher model is updated by student model, the biased representation would be accumulated in the mutual learning framework. This issue would affect the generalization ability of model and undermine the converging speed. Conversely, a tolerant selective strategy would guarantee the quantity of pseudo labels, but more noises are introduced in mutual learning frameworks. The accumulated error would also generate adverse effects on detectors. Based on the above analysis, we believe
that the strategy of selecting pseudo labels is the dominant reason for limiting the performance in semi-supervised single-stage object detectors.

4 Method

Previous semi-supervised object detection methods use the empirical strategy of selecting pseudo labels according to the ablation experiments [35,19]. By analyzing the distribution gap between single- and two-stage detectors, we believe that the quality and quantity of pseudo labels should be considered equally when choosing the strategy in semi-supervised learning. To this end, we propose a dynamic self-adaptive threshold (DSAT) strategy, which can dynamically adjust the strategy based on model itself and achieve an optimal trade-off between quality and quantity of pseudo labels. Besides, to assess the regression quality of pseudo labels, we develop a module named NMS-UNC that could compute uncertainty of them based on NMS. Fig. 2 shows the overall pipeline of the proposed semi-supervised learning for single-stage detectors (S4OD). As shown in Fig. 2, we feed the dense predictions generated from teacher model into the NMS-UNC module to get the sparse pseudo labels including classification scores, boxes and uncertainty. An uncertainty threshold is used to preserve boxes with lower uncertainty. These preserve boxes are regression targets for unlabeled data. Another path use the DSAT to select boxes with higher classification scores. These boxes are assigned with one-hot labels for classification of unlabeled data.
Fig. 3. (a) Precision, Recall and F1-score distribution of classification on FCOS. (b) Sampling ratio and corresponding mAP of w/ and w/o DSAT on FCOS. Compared with original FCOS, the proposed DSAT relieves the sampling imbalance (orange points) and improves the performance (blue points).

4.1 Dynamic Self-Adaptive Threshold Strategy

Since the selection of pseudo labels is of vital importance for semi-supervised learning, we propose a dynamic self-adaptive threshold strategy which can adjust the threshold automatically based on teacher model itself during the whole training process. To ensure the quality and quantity of pseudo labels, the strategy should consider both precision and recall of pseudo labels. However, precision and recall are in conflict, it’s not possible to gain a higher precision with a higher recall. To reasonably make a trade-off on precision and recall, we heuristically choose the F1-score as the index to design the threshold strategy, because the F1-score denotes the harmonic mean of precision and recall. F1-score can be formulated by Eq. (3):

\[
F1\text{-score} = \frac{2 \times Pre \times Rec}{Pre + Rec}
\]

\[P_{re}\] and \[Rec\] denote the precision and recall of the prediction. For an object detection model, we first pick a set of confidence scores (range from 0.05 to 0.95 at step 0.05). For each confidence score, we preserve all pseudo labels whose classification scores are larger than it. Then the corresponding precision, recall and F1-score on the validation set can be computed with a pre-defined IoU threshold (0.5 is used in this paper). Finally, as shown in Fig 3(a), the curve of precision, recall and F1-score can be obtained by connecting the corresponding points.

To choose a proper pseudo label selecting strategy which can take both quality and quantity of pseudo labels into consideration, it should be set at the peak of the F1-score curve to achieve a relatively high quality without losing too many pseudo labels for student model. Compared with choosing a strict strategy to
Fig. 4. (a) The confidence score corresponding to the F1-score peak during training. (b) The distribution between the standard deviation of redundant boxes and IoU. The fit line (in red) shows the relation between them.

ensure the precision of pseudo labels, picking confidence score with respect to the F1-score peak could preserve more objects so that the imbalance of positive and negative samples can be alleviated (as shown in Fig.3(b)). Using the F1-score curve as the reference to choose strategy only depends on model performance, which is self-adaptive. Meanwhile, it eases the implementation greatly without the need of designing a over-complicated strategy.

As the model iterates, the performance of the model is increased accordingly. In this case, the confidence corresponding to the F1-score peak is dynamic instead of immutable during the training process (as shown in Fig.4(a)). As a result, using the fixed strategy would ignore more objects at the early stage and slow down the converging speed, which can aggravate the effect on the model’s generalization ability. To this end, we further design a dynamic self-adaptive threshold strategy. Specifically, the F1-score curve of the teacher model can be obtained at the validation stage during training. Then we choose the confidence score corresponding to the maximum F1-score value as the filtering threshold of pseudo labels for the next training stage. Through the dynamic self-adaptive threshold strategy, the semi-supervised framework could build the most efficient information flow from the teacher model to the student model, which not only speeds up the training process but also promotes the performance.

4.2 NMS-UNC

Another factor that affects the quality of pseudo-labels is regression. Different from classification, it is hard to assess the regression quality of pseudo labels in a continual domain. Inspired by the discretized continuous regression methods [14,38], we propose NMS-UNC, an assessment method of box regression based on NMS (non-maximum suppression) in single-stage detectors.
Algorithm 1 NMS-UNC

Input: Dense prediction $(B_{box}, B_{score})$, IoU threshold $\delta$, Score threshold $\mu$

Output: Reserved prediction $(P_{box}, P_{score}, P_{unc})$

1 $P_{box} \leftarrow \emptyset$, $P_{score} \leftarrow \emptyset$, $P_{unc} \leftarrow \emptyset$
2 $B_{score} \leftarrow B_{score}[B_{score} \geq \mu]$, $B_{box} \leftarrow B_{box}[B_{score} \geq \mu]$

3 while $B_{box}$ is not empty do
4 $I \leftarrow \text{Max}(B_{score})$
5 $B_{m} \leftarrow B_{box}[I]$, $S_{m} \leftarrow B_{score}[I]$
6 $B_{c} \leftarrow \emptyset$
7 for $n = 0$ to $\text{Size}(B_{box})$ do
8 $\text{iou} \leftarrow \text{IoU}(B_{box}[n], B_{m})$
9 if $\text{iou} \geq \delta$ then
10 $B_{c} \leftarrow B_{c} \cup B_{box}[n]$
11 del $B_{box}[n]$
12 end
13 end
14 if $\text{Size}(B_{c}) > 1$ then
15 $\text{unc} \leftarrow \text{std}(B_{c})$
16 $P_{box} \leftarrow P_{box} \cup B_{m}$, $P_{score} \leftarrow P_{score} \cup S_{m}$, $P_{unc} \leftarrow P_{unc} \cup \text{unc}$
17 end
18 end

For the dense predictions of the single-stage object detectors, we find that the predicted dense bounding boxes are relatively compact for the simple objects (whose boundaries are clear) while scattered for the hard objects. Based on this observation, we believe that the standard deviation of these redundant bounding boxes that belong to one object can reflect the regression uncertainty of this object. As the statistics shown in Fig.4(b), the bounding box with a larger standard deviation generally has a lower IoU with the ground-truth, and the fit line can indicate the relation between the proposed standard deviation with the IoU. Therefore, the standard deviation could be used to assess the regression quality. As a result, we design the NMS-UNC module which can compute the uncertainty (standard deviation) of the reserved bounding boxes during the NMS. Algorithm 1 elaborates the detail process. Consistent with the original NMS, NMS-UNC first filters out boxes whose classification scores below the score threshold $\mu$. Then, for the maximum score bounding box $B_{m}$ in each iteration, NMS-UNC gathers the boxes $B_{c}$ whose IoU with the maximum score box $B_{m}$ are larger than the IoU threshold $\delta$. These boxes are regarded as the redundant boxes in the original NMS module and should be removed. Before these boxes are removed, we compute the standard deviation of these boxes $B_{c}$ to obtain the uncertainty of $B_{m}$. This process can be formulated by Eq. (4):

$$\text{uncertainty} = \frac{1}{4} \sum_{i=0}^{1} \sum_{j=0}^{1} \frac{\text{std}(B_{c})[i + j]}{\tilde{w}} + \tilde{h})/4$$  (4)
where \( \text{std} \in \mathbb{R}^4 \) denotes the standard deviation of redundant boxes \( B_c \). Notably, \( i \) decides the top-left point or bottom-right point of the box and \( j \) indicates the coordinate axis \( x \) or \( y \). \( \bar{w} \) and \( \bar{h} \) denote the mean width and mean height of boxes.

After the NMS-UNC post-processing module, we can get the classification score and the regression uncertainty for each sparse pseudo label. To guarantee the quality of pseudo labels, two thresholds are set to constrain the classification score and the regression result separately. By introducing the uncertainty of pseudo labels, the unsupervised loss \( L_{\text{unsup}} \) is then formulated by Eq. (5):

\[
L_{\text{unsup}} = \frac{1}{N_{\text{pos}}} \sum_{x,y} L_{\text{cls}}(p_{x,y}, c_{x,y}^*) + \frac{1}{N_{\text{pos}}} \sum_{i} L_{\text{reg}}(t_i, \hat{b}_i) \tag{5}
\]

where \( N_{\text{pos}} \) and \( c^* \) are the number of positive samples and classification target according to the pseudo classification label whose scores are higher than \( \sigma_{\text{cls}} \). \( \hat{b} \) is the regression targets whose uncertainty is smaller than regression threshold \( \sigma_{\text{unc}} \). \( N_{\text{pos}}^b \) is the number of positive samples assigned from these regression targets \( \hat{b} \).

5 Experimental Results

5.1 Implementation Details

**Dataset and Metrics.** All experiments are conducted on the MS-COCO [17] dataset including 80 object categories and 118k images. For semi-supervised training, we follow the setting from Xu et al. [35] and split the COCO-train2017 set (all labeled) into multiple datasets with different labeled percentages: 1%, 5% and 10%. Besides, we use the COCO-unlabel2017 to conduct incremental experiments including 123k images. The COCO-val2017 containing 5,000 images is used to evaluate the mean average precision (mAP) metrics.

**Training Details.** We implement the anchor-free single-stage detector (FCOS [29]) and the anchor-based single-stage detector (RetinaNet [16]) based on MMdetection [5]. Different from classification and regression branches in anchor-based detectors, the dense output from FCOS includes another centerness branch. We use the classification score rescaled by centerness as the final score and compare the score with the threshold to select high-quality pseudo labels. ResNet-50 [10] with FPN [15] is used and the student model is trained by the SGD optimizer for 180K iterations with 0.9 momentum and 0.0001 weight decay. The learning rate is set to 0.01 with a warm-up strategy and decays by 0.1 at iteration 120K and 160K. For teacher model, the EMA update rate is set to 0.999. The same data-augmentation method used by Xu et al. [35] is applied in our experiments. Batchsize is 32 and 8 Nvidia Tesla P40 GPUs are used for training. Unless otherwise specified, all experiments are conducted as above.

For semi-supervised baseline, we set the fixed classification threshold \( \sigma_{\text{cls}} \) to 0.5 as it achieves the best performance in Table 4. For dynamic threshold...
Table 1. Results on different labeled percentages.

| Model       | Supervised     | 1%       | 5%       | 10%      |
|-------------|----------------|----------|----------|----------|
| Fully Baseline | 10.0          | 20.92    | 26.94    |
| Faster-RCNN  | STAC[26]       | 13.97(+3.97) | 24.38(+3.46) | 28.64(+1.70) |
| Semi        | Unbiased[19]   | 20.75(+10.75) | 28.27(+7.35) | 31.50(+4.56) |
|             | SoftTeacher[35] | 20.46(+10.46) | 30.74(+9.82) | 34.04(+7.10) |
| FCOS[29]    | Fully Baseline | 13.6     | 23.5     | 28.1     |
| Semi        | Ours           | 23.7(+10.1) | 32.3(+8.8) | 35.0(+6.9) |
| RetinaNet[16] | Fully Baseline | 12.2     | 22.7     | 27.4     |
| Semi        | Ours           | 20.1(+7.9) | 30.0(+7.3) | 32.9(+5.5) |

Table 2. Experiments of using all labeled data of COCO-train2017 and the extra unlabeled data of COCO-unlabel2017.

| Labeled    | Unlabeled | mAP |
|------------|-----------|-----|
| Baseline   | COCO-train2017 | -   | 37.9 |
| Ours       | COCO-train2017 | COCO-unlabel2017 | 42.8(+4.9) |

setting, we evaluate the F1-score at each 4,000 iterations and update the selecting strategy according to the latest F1-score peak. As for uncertainty threshold, we use $\sigma_{unc} = 0.08$ according to the ablation study result in Table 5. The original score threshold $\mu$ is 0.05 in NMS, which would preserve many noise bounding boxes for uncertainty computing in NMS-UNC. Therefore, we increase the score threshold $\mu$ to 0.4 for preserving the valuable bounding boxes to assess the regression quality. For other hyper-parameters in FCOS and RetinaNet, we follow the default setting in MMdetection.

**Inference Details.** During the semi-supervised training, the teacher model always performs better than student model. As a result, we choose the teacher model as the final model and use the same sizes of input images as training. Because only the teacher model is used to inference, the semi-supervised learning would not increase the inference time. Besides, to make a fair comparison, NMS score threshold $\mu$ is set to 0.05 in inference and evaluation.

5.2 Results Comparison

To verify the effectiveness of semi-supervised learning in single-stage detectors, we conduct elaborate experiments and report the results in Table 1 and Table 2. Table 1 shows that the proposed semi-supervised learning has achieved a significant improvement compared with fully-supervised learning. The semi-supervised setting outperforms the fully-supervised training by more than +5.5%
Table 3. Ablation study on the effectiveness of the proposed modules.

| Method          | DSAT | NMS-UNC | AP  | AP_{50} | AP_{75} |
|-----------------|------|---------|-----|---------|---------|
| Fully-supervised| 23.5 | 39.5    | 24.0|         |         |
|                 |      |         | 29.8| 47.9    | 31.3    |
| Semi-supervised | ✓    | ✓       | 31.3| 49.1    | 33.0    |
|                 | ✓    | ✓       | 31.6| 49.2    | 33.3    |
|                 | ✓    | ✓       | 32.3| 50.1    | 34.0    |

Table 4. Ablation study on the fixed classification threshold $\sigma_{cls}$.

| $\sigma_{cls}$ | AP  | AP_{50} | AP_{75} |
|-----------------|-----|---------|---------|
| 0.4             | 29.3| 47.2    | 30.6    |
| 0.5             | **29.8**| **47.9**| **31.3**|
| 0.6             | 29.1| 46.2    | 30.5    |
| 0.7             | 24.2| 39.0    | 25.2    |
| DSAT (ours)     | **31.6** (+1.8)| **49.2** (+1.3)| **33.3** (+2.0) |

AP, +10.1\% at most. These improvements convince that by automatically adjusting the selective strategy and using the proposed NMS-UNC to choose regression targets, single-stage object detectors can achieve excellent improvements as well.

Specifically, for anchor-free object detector (FCOS), the proposed method consistently boosts the performances by +10.1\%, +8.8\% and +6.9\% AP in 1\%, 5\% and 10\% labeled data setting correspondingly. For anchor-based single-stage detector (RetinaNet), the improvements are +7.9\%, +7.3\% and +5.5\% on AP.

In addition, we conduct incremental experiments on all labeled data of COCO-train2017 set on FCOS. As shown in Table 2 when using all data of COCO-train2017 set, the fully-supervised learning achieves 37.9\% AP. By leveraging the unlabeled data of COCO-unlabel2017 additionally, our S4OD method reaches 42.8\% on AP, which is 4.9\% higher than the baseline.

5.3 Ablation Studies

**Proposed Modules.** We design ablation experiments on FCOS to analyze the effectiveness of the proposed DSAT strategy and NMS-UNC module. Table 3 shows that both the DSAT strategy and the NMS-UNC module outperform the semi-supervised method in all different labeled percentages. Most importantly, the model achieves the best results when combining the two modules, which means that the proposed modules are orthogonal.

**Fixed Classification Threshold.** For the fixed classification threshold, we conduct different classification threshold settings on FCOS to find the best-performed classification threshold. As shown in Table 4 when the classification
Table 5. Ablation study on the uncertainty threshold $\sigma_{unc}$.

| $\sigma_{unc}$ | AP  | AP_{0.5} | AP_{0.75} |
|----------------|-----|----------|-----------|
| 0.05           | 31.0| 48.3     | 32.9      |
| 0.08           | **31.3**| **49.1**| **33.0** |
| 0.10           | 31.0| 48.8     | 32.9      |

threshold is 0.5, the models achieve the best performance. Notably, compared with the lower thresholds, the models are more sensitive to the higher thresholds. When increasing the threshold to 0.7, the performance drops sharply. These results support the view of fewer pseudo labels lead to a more severe sampling imbalance of positive and negative. It would waste expensive time and computing cost to find a suitable classification threshold by ablation experiments, which inspires us to design a dynamic self-adaptive threshold strategy. The bottom row lists our DSAT results. Compared with the best fixed threshold setting, the proposed DSAT achieves +1.8% improvement on AP.

**Uncertainty Threshold.** By introducing the proposed NMS-UNC module, we can calculate the regression uncertainty for each pseudo label. A well-chosen uncertainty threshold can meet the demand of filtering out the pseudo labels with higher uncertainty. Table 5 shows that when the threshold is 0.08, the model achieves the best performance. The similar conclusion can be drawn from Fig.4(b), the scatters with higher IoU become more sparser when the standard deviation (uncertainty) is higher than 0.08. For uncertainty threshold 0.05, it would abandon a large number of well-regressed bounding boxes and affect the generalization of models. As a result, we use $\sigma_{unc} = 0.08$ in our experiments for optimal performances.

**Score Threshold in NMS-UNC.** In the original NMS, a score threshold $\mu$ is used to select high quality boxes and reduce the computation. To ensure the recall of detectors, $\mu$ usually sets to 0.05. In this paper, to assess the regression quality for selecting pseudo labels, we proposed the NMS-UNC module based on NMS, which computes the standard deviation of these redundant boxes as its uncertainty before removing them. To avoid the affect caused by the low-quality boxes in computing uncertainty, the score threshold $\mu$ should be increased properly. As a result, we design ablation experiments on the score threshold $\mu$ and list the result in Table 6. From Table 6, the NMS-UNC module achieves the best performance when the score threshold $\mu$ is 0.4. Besides, $\mu = 0.4$ is more efficient in computing compared with $\mu = 0.3$, because $\mu = 0.3$ would involve more boxes.

### 5.4 Discussion

**Time Comparison.** In this section, we make a time analysis on the proposed DSAT and NMS-UNC. The increased time cost of DSAT strategy comes from the F1-score computing. As shown in Table 7, the time cost of F1-score computing is
Table 6. Ablation study on the score threshold $\mu$ in NMS-UNC.

| $\mu$ | AP | $AP_{0.5}$ | $AP_{0.75}$ |
|-------|----|------------|-------------|
| 0.3   | 31.0 | 48.4 | 33.0 |
| 0.4   | **31.3** | **49.1** | **33.0** |
| 0.5   | 30.7 | 48.0 | 32.2 |

Table 7. Running time analyze of the proposed module.

| DSAT | NMS-UNC | sec/iter | total time(sec) |
|------|---------|----------|-----------------|
| ✓    | 2.470   | 444600   |                 |
| ✓    | 2.520   | 453600(+2.02%)|               |
| ✓    | 2.607   | 469260(+5.55%)|               |

negligible because only 2.02% training time increased. NMS-UNC module needs to compute the box uncertainty based on the original NMS module. We integrate the uncertainty computing process with the original CPU version NMS instead of the CUDA version NMS for simplified implementation, which would slow down the training. However, this process can be accelerated by utilizing CUDA Core if necessary. The increased time is acceptable compared with the whole training process. Moreover, the uncertainty is only computed in training, there is no additional time cost for inference and evaluation.

Application Value. The proposed DSAT module is heuristic, because it has no hypothesis to the model and data, which means that the proposed strategy could help the semi-supervised learning to be applied in most engineering object detection tasks. Besides, although the proper filter strategy can be determined by conducting sophisticated ablation experiments, the performance of the fixed classification threshold is still lower than that of the DSAT strategy (as shown in the second and fourth rows of Table 3). In conclusion, our DSAT strategy outperforms the fixed threshold strategy prominently not only in the implementation cost but also in performance.

6 Conclusions

In this work, we point out that the improper strategy of selecting pseudo labels would intensify the imbalance of positive and negative samples of single-stage detectors in semi-supervised learning. As a result, how to choose a proper strategy is critical to semi-supervised learning for single-stage object detectors. To guarantee both the quality and quantity of pseudo labels, we propose a dynamic self-adaptive threshold strategy, which could adjust the selecting strategy according to the model performance automatically. Besides, we also design a measurement method of box regression quality for single-stage detectors. Extensive experiments in the MS-COCO dataset convince that the proposed method
S4OD: Semi-Supervised learning for Single-Stage Object Detection. achieves state-of-the-art performance in semi-supervised single-stage object detection.
References

1. Assran, M., Caron, M., Misra, I., Bojanowski, P., Joulin, A., Ballas, N., Rabbage, M.: Semi-supervised learning of visual features by non-parametrically predicting view assignments with support samples. In: ICCV (2021)
2. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.: Mixmatch: A holistic approach to semi-supervised learning. arXiv preprint arXiv:1905.02249 (2019)
3. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: CVPR (2018)
4. Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Shi, J., Ouyang, W., et al.: Hybrid task cascade for instance segmentation. In: CVPR (2019)
5. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C.C., Lin, D.: MMDetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
6. Chen, X., Yuan, Y., Zeng, G., Wang, J.: Semi-supervised semantic segmentation with cross pseudo supervision. In: CVPR (2021)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021)
9. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV (2017)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
11. Jeong, J., Lee, S., Kim, J., Kwak, N.: Consistency-based semi-supervised learning for object detection. In: NeurIPS (2019)
12. Jeong, J., Verma, V., Hyun, M., Kannala, J., Kwak, N.: Interpolation-based semi-supervised learning for object detection. In: CVPR (2021)
13. Jiang, B., Zhang, Z., Lin, D., Tang, J., Luo, B.: Semi-supervised learning with graph learning-convolutional networks. In: CVPR (2019)
14. Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., Tang, J., Yang, J.: Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) NeurIPS (2020)
15. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR (2017)
16. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV (2017)
17. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
18. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: ECCV (2016)
19. Liu, Y.C., Ma, C.Y., He, Z., Kuo, C.W., Chen, K., Zhang, P., Wu, B., Kira, Z., Vajda, P.: Unbiased teacher for semi-supervised object detection. arXiv preprint arXiv:2102.09480 (2021)
20. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al.: Swin transformer v2: Scaling up capacity and resolution. arXiv preprint arXiv:2111.09883 (2021)
21. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV (2021)
22. Pham, H., Dai, Z., Xie, Q., Le, Q.V.: Meta pseudo labels. In: CVPR (2021)
23. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
24. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: NeurIPS (2015)
25. Sohn, K., Berthelot, D., Li, C.L., Zhang, Z., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H., Raffel, C.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. arXiv preprint arXiv:2001.07685 (2020)
26. Sohn, K., Zhang, Z., Li, C.L., Zhang, H., Lee, C.Y., Pfister, T.: A simple semi-supervised learning framework for object detection. arXiv preprint arXiv:2005.04757 (2020)
27. Tang, Y., Chen, W., Luo, Y., Zhang, Y.: Humble teachers teach better students for semi-supervised object detection. In: CVPR (2021)
28. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: NeurIPS (2017)
29. Tian, Z., Shen, C., Chen, H., He, T.: Fcos: Fully convolutional one-stage object detection. In: ICCV (2019)
30. Wang, C.Y., Yeh, I.H., Liao, H.Y.M.: You only learn one representation: Unified network for multiple tasks. arXiv preprint arXiv:2105.04206 (2021)
31. Wang, Z., Li, Y.L., Guo, Y., Wang, S.: Combating noise: Semi-supervised learning by region uncertainty quantification. In: NeurIPS (2021)
32. Wang, Z., Li, Y., Guo, Y., Fang, L., Wang, S.: Data-uncertainty guided multi-phase learning for semi-supervised object detection. In: CVPR (2021)
33. Wei, C., Sohn, K., Mellina, C., Yuille, A., Yang, F.: Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In: CVPR (2021)
34. Xie, Q., Dai, Z., Hovy, E., Luong, M.T., Le, Q.V.: Unsupervised data augmentation for consistency training. arXiv preprint arXiv:1904.12848 (2019)
35. Xu, M., Zhang, Z., Hu, H., Wang, J., Wang, L., Wei, F., Bai, X., Liu, Z.: End-to-end semi-supervised object detection with soft teacher. arXiv preprint arXiv:2106.09018 (2021)
36. Yang, Q., Wei, X., Wang, B., Hua, X.S., Zhang, L.: Interactive self-training with mean teachers for semi-supervised object detection. In: CVPR (2021)
37. Zhai, X., Oliver, A., Kolesnikov, A., Beyer, L.: S4i: Self-supervised semi-supervised learning. In: ICCV (2019)
38. Zheng, Z., Ye, R., Wang, P., Ren, D., Zuo, W., Hou, Q., Cheng, M.: Localization distillation for dense object detection. In: CVPR (2022)
39. Zhou, Q., Yu, C., Wang, Z., Qian, Q., Li, H.: Instant-teaching: An end-to-end semi-supervised object detection framework. In: CVPR (2021)
40. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)