Scale-Equivalent Distillation for Semi-Supervised Object Detection

Qiushan Guo[1], Yao Mu[1], Jianyu Chen[2], Tianqi Wang [1], Yizhou Yu[1], Ping Luo[1]
[1]The University of Hong Kong [2]Tsinghua University
{qsguo,ymu,tqwang,yzyu,pluo}@cs.hku.hk jianyuchen@tsinghua.edu.cn

Abstract

Recent Semi-Supervised Object Detection (SS-OD) methods are mainly based on self-training, i.e., generating hard pseudo-labels by a teacher model on unlabeled data as supervisory signals. Although they achieved certain success, the limited labeled data in semi-supervised learning scales up the challenges of object detection. We analyze the challenges these methods meet with the empirical experiment results. We find that the massive False Negative samples and inferior localization precision lack consideration. Besides, the large variance of object sizes and class imbalance (i.e., the extreme ratio between background and object) hinder the performance of prior arts. Further, we overcome these challenges by introducing a novel approach, Scale-Equivalent Distillation (SED), which is a simple yet effective end-to-end knowledge distillation framework robust to large object size variance and class imbalance. SED has several appealing benefits compared to the previous works. (1) SED imposes a consistency regularization to handle the large scale variance problem. (2) SED alleviates the noise problem from the False Negative samples and inferior localization precision. (3) A re-weighting strategy can implicitly screen the potential foreground regions of the unlabeled data to reduce the effect of class imbalance. Extensive experiments show that SED consistently outperforms the recent state-of-the-art methods on different datasets with significant margins. For example, it surpasses the supervised counterpart by more than 10 mAP when using 5% and 10% labeled data on MS-COCO.

1. Introduction

Deep neural networks achieve strong results under the supervised learning framework driven by large-scale datasets, such as ImageNet [5] (about 1.28 million labeled images). However, different from classification, object detection further involves locating objects with a bounding box. Therefore, the annotation for object detection is much more expensive, leading to labeled data remaining scarcely related to classification. Recently, Semi-Supervised Learning (SSL) for classification has received much attention [2, 29, 33, 35], whose results are comparable to the fully supervised model on ImageNet. However, Semi-supervised Object Detection (SS-OD) is more challenging than SSL on ImageNet classification. Recent SS-OD methods improve the performance by leveraging both the limited labeled data and the massive unlabeled data, but they suffer from the large variance of object sizes, massive False Negative samples and class imbalance problem, as illustrated in Fig. 1.

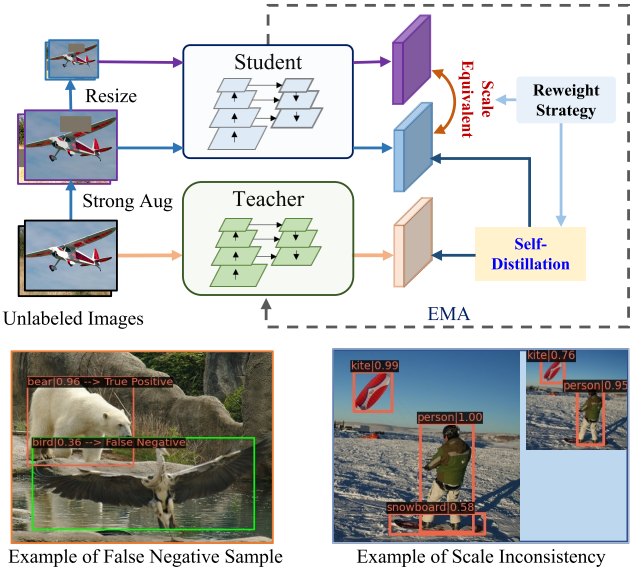


Figure 1. The overall framework of SED. Our model improves the scale equivalence, which is critical for object detectors, by regularizing the consistency between different-sized images. Furthermore, the inherent False Negative sample noise is alleviated by self-distillation. A re-weighting strategy is adopted to solve the severe class imbalance problem. The bird in the left example is a False Negative sample when the threshold is set to 0.7. The right example shows the scale inconsistency of different-sized images.

The scale of objects varies in a small range for the ImageNet classification model, whereas the scale variation of MS-COCO dataset [18] is large across object instances for the detector. As shown in Fig. 2a, the standard deviation of

the scale of instances in MS-COCO is 188.4 pixels, while that of ImageNet is 56.7 pixels (the square root of area). A detector is supposed to be scale consistent to object instances, which means that the predictions of an image in different sizes should be equivalent [27, 28]. However, the scale consistency has not been considered by the prior arts [19, 30, 36, 39] in SS-OD. We observe a discrepancy in the objectness score, as indicated in Fig. 2b. The ratio of foreground anchor to background anchor increases as the score distance becomes large, which implies that the model detects an object instance while is blind to the instance in a different size. This inconsistency is typically alleviated by the multi-scale inference ensemble, which increases the computational cost and requires complicated operations to fuse the results.

Besides, the performance of recent SS-OD methods [19, 29] is moderate in the high-data scenario as a consequence of the False Negative object instance and inferior localization precision. As illustrated in Fig. 2c, the recall drops to 0.1 and 0.3 separately when IoU is set to 0.5 and 0.9, which indicates that most foreground instances are False Negative samples. The precision at IoU = 0.9 is less than 0.2, showing that the location of bounding boxes is not accurate enough. The False Negative object instances below the hard threshold cause a recognition inconsistency.

Another obstacle is that the foreground and background samples are highly imbalanced. The ratio of the foreground to background sample is about 1:25,000 for RetinaNet [17]. Due to the class imbalance problem, treating all regions equally [32] leads to the background samples contributing significantly to the gradient, as illustrated in Fig. 4. Identifying foreground regions from the unlabeled data with the overwhelming background regions is challenging.

To overcome the challenges motioned above, we propose Scale-Equivalent Distillation (SED), a simple yet effective end-to-end semi-supervised learning framework for object detection. Since scale is an essential factor of the low-dimensional semantic manifolds, we design a scale consistency regularization across the prediction in different levels as a solution to the large object size variance. Moreover, as the noise from hard pseudo-label has detrimental effects on the recognition consistency, a self-distillation method is proposed to improve generalization performance without increasing the learnable parameters. Due to the class imbalance problem, the overwhelming background samples diminish the effect of our method. We implement a re-weighting strategy to focus on the inconsistency among outputs in different levels and the discordance between the teacher and student detector. As a result, our re-weighting approach avoids selecting the potential foreground regions from the unlabeled data explicitly.

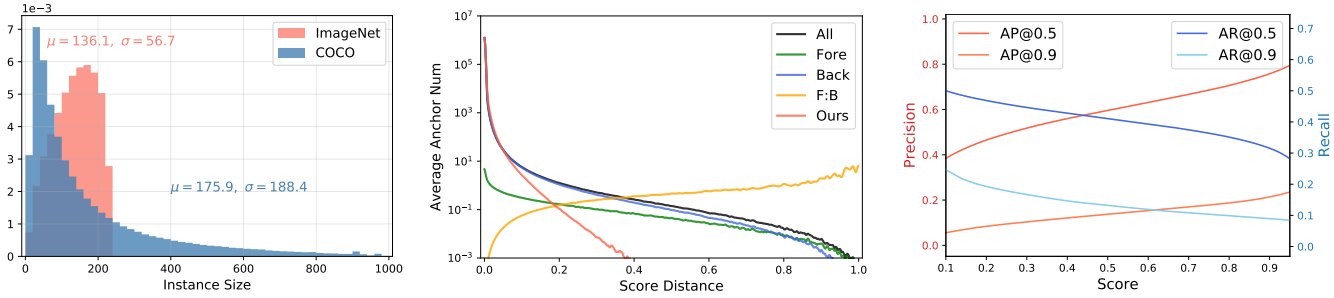
To evaluate the validation of SED, we conduct extensive experiments on benchmarks for object detection, Pascal VOC [7] and MS-COCO [18]. Our method surpasses the supervised counterpart by more than 10 mAP when using 5% and 10% labeled data on MS-COCO. Moreover, our method is tested with both one-stage and two-stage detector based on single feature map and feature pyramid.

Our contributions are listed as follows: (1) SED imposes a scale consistency regularization to overcome the large scale variance challenge. (2) SED alleviates the noise problem which arises from the False Negative samples and inaccurate bounding box regression. (3) A re-weighting strategy can implicitly screen the potential foreground regions from unlabeled data to reduce the effect of class imbalance.

2. Related Works

Self-Training. Self-training methods first train a teacher model with the labeled dataset and then generate pseudo-labels for the unlabeled dataset. Finally, the student model is optimized with both the labeled data and pseudo-labeled data jointly. For the classification task, Self-training methods [1, 2, 29, 33] perform well. However, Semi-Supervised Object Detection is more challenging than Semi-Supervised Image Classification on the balanced dataset. Some works [19, 39] contribute to alleviating the noise problem brought by pseudo-label. Those methods attach additional modules on the two-stage detector to overcome the heavy overfitting problem on the foreground and background classification and refine the hard pseudo-label by ensemble methods. Nevertheless, methods based on hard pseudo-label have an inherent defect that False Negative object instances influence the consistency of recognition, especially whose scores are near the hard threshold. Humble Teacher [32] adopts soft pseudo-labels to avoid the recognition inconsistency but treat all the regions equally. Due to the extreme imbalance of foreground and background, the contribution of gradients from the two kinds of regions is quite different. UBT [19] adopts Focal Loss to alleviate the problem. Unlike the existing works, our method generates soft pseudo-labels for unlabeled data in an online manner, and the re-weighting strategy automatically screens the potential foreground regions of the unlabeled data.

Consistency Regularization. Consistency-based Semi-supervised learning uses unlabeled data to stabilize the predictions under input or weight perturbations. For instance, two different views of the same image are supposed to have similar output. This class of methods [20, 26, 33] does not generate pseudo-label but constrains the discrepancy between the outputs, which is known to help smooth the manifold [21]. For SS-OD, CSD [14] applies simple horizontal flip consistency regularization to train a detector to be robust to flip perturbations. The consistency loss fine-tunes the location of the predicted boxes but ignores the object scale perturbations, which are more common in datasets. In MS-COCO [18] detection dataset, the scale of the smallest and



(a) Instance Size Distribution (b) Score Distance Distribution (c) AP and AR *w.r.t* Score

Figure 2. (a) For the COCO dataset, all the images are resized such that the short edge has 800 pixels while the long edge has less than 1333 pixels. For the ImageNet dataset, all the images are resized to 224×224 to calculate the statistics. The scale of object is represented as the square root of the area. We discuss the typical training input size for ImageNet classification and COCO detection tasks. (b) All the scores are predicted on COCO *minival* dataset by the RetinaNet detector with FPN and ResNet 50 backbone, which is trained with 10% COCO data. The score distance is the absolute difference between the predictions of the image in different sizes. The Y-axis is the average number of anchors per image. (c) We predict pseudo-label on the rest of COCO training data with a converged Faster-RCNN detector (with FPN and ResNet50 backbone), trained with 10% COCO data. The low average recall and precision show that hard pseudo-label incur more noise with False Negative samples.

largest 10% of object instances is 0.024 and 0.472, respectively. Our method regularizes the predictions of different sizes to solve the large-scale variation. Furthermore, self-distillation [8, 10, 38] benefits from the high-quality prediction of EMA teacher [33], and can be viewed as consistency regularization from the perspective of soft targets.

Pre-Training. In recent years, it has been a paradigm that pre-train backbone on a large-scale dataset, such as ImageNet [5] or JFT [31], and fine-tune the model on the target dataset, which contains less training data. Large-scale dataset pre-training speeds up converge and helps improve generalization in the small data scenario [12, 40], which is an extreme of semi-supervised learning. SimCLR [4] and MOCO [11] have been shown to build universal representation, which helps achieve a state-of-the-art result in the scenario of semi-supervised learning classification with 10% ImageNet labeled data. In this paper, we fine-tune with ImageNet pre-trained backbone as default for faster convergence and better results when we enter the low-data regime.

3. Scale-Equivalent Distillation

Problem Definition. Semi-supervised learning is halfway between supervised and unsupervised learning. More precisely, our model is trained with a labeled set $D_\text{s} = \{x_i^\text{s}, y_i^\text{s}\}_{i=1}^{N_\text{s}}$ and an unlabeled set $D_\text{u} = \{x_i^\text{u}\}_{i=1}^{N_\text{u}}$, where x is image, N_s and N_u are the number of labeled and unlabeled images. For each supervised image x_i^s, the annotation y_i^s is composed of both the location and category of the bounding boxes in image.

Overview. During training, Scale-Equivalent Distillation consists of two branches, the supervised and unsupervised branch, as illustrated in Fig. 3. The supervised branch is trained by following the normal procedure, like [17, 24]. The unsupervised branch is under a teacher-student framework, in which the teacher is implemented as an exponential moving average of the student. SED aims to predict consistently for the scale variants of input. In practice, the student processes the strongly augmented unlabeled images and resized images. The weakly augmented images are fed into the teacher network to predict soft pseudo-label. The scale consistency loss constrains the outputs of different-sized images. Meanwhile, the soft pseudo-label is set as the target of the strongly augmented images. The final loss is the weighted sum of the supervised and unsupervised loss,

$$L = L_\text{supervised} + \frac{n_\text{u}}{n_\text{s}}(\lambda_\text{s} L_\text{scale} + \lambda_d L_\text{distill}), \qquad (1)$$

where n_u, n_s are the batch size of unlabeled data and labeled data. L_scale and L_distill are Scale Consistency Loss and Self-Distillation Loss. For two-stage detectors, the unsupervised losses are applied to both the RPN and RoI head.

3.1. Scale Consistency Regularization

Recognizing objects in different scales is a fundamental challenge in computer vision. Scale Consistency Regularization is proposed to optimize the detector to predict smoothly and consistently in scale dimension. Typically, mainstream detectors under the feature pyramid network (FPN) framework outperform a single feature map counterpart, as the multi-scale feature representations are semantically strong. Therefore, we take an example for a single-stage detector with FPN to illustrate our method. Scale Consistency Regularization can be extended to the two-stage detectors and single feature map detectors.

As indicated in Fig. 3, scale consistency loss regularizes predictions from images in different scales. To be more specific, the output class probability and bounding box regression of the f-th feature level, r-th row, c-th column and d-th anchor box are denoted as $P^{f,r,c,d}(X)$ and $R^{f,r,c,d}(X)$. Considering the memory and calculational cost, the resized

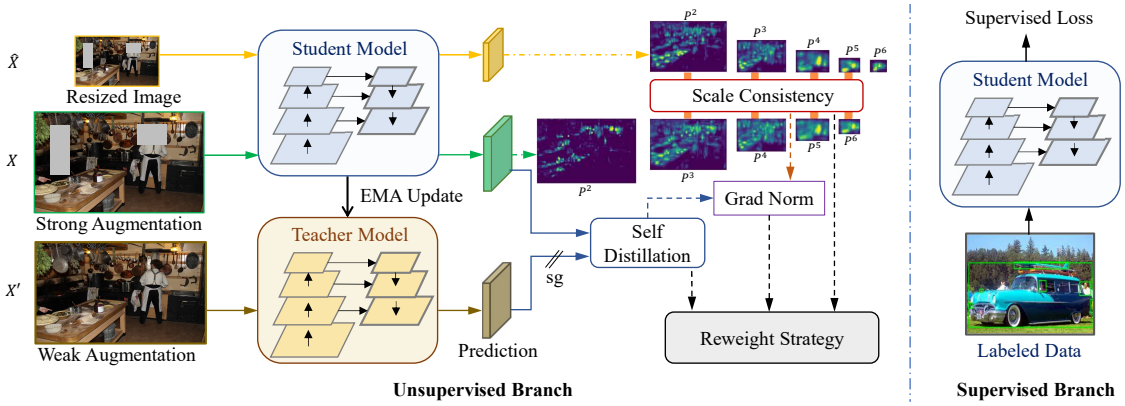


Figure 3. Details of our method. We take an example of a detector with FPN [16] to illustrate our method. P^2-P^6 are the prediction details.The supervised branch shares the Student Model with the unsupervised branch. *sg* means the prediction of the Teacher Model is not optimized by the gradient. For Scale Consistency Regularization, the loss constrains the predictions from different levels.

image is downsampled to $\frac{1}{2^s}$ original size. Towards handling the large scale variation, the s is uniformly selected from $\{1, 2, ..., S\}$, which also matches the sizes of feature maps in FPN and the label assignment rules. The resized image $\hat{X}$ and the original image X are supposed to be predicted consistently for the corresponding levels. Precisely, the scale consistency loss is defined as

$$\begin{aligned} L^f_{\text{scale}} =& D_{\text{KL}}(\text{sg}(P^f(X)), P^{f'}(\hat{X})) \\ &+ D_{\text{KL}}(\text{sg}(P^{f'}(\hat{X})), P^f(X)) \\ &+ ||R^f(X) - R^{f'}(\hat{X})||_2, \end{aligned} \tag{2}$$

where f' equals $f - s$ and sg is stop-gradient operator. For simplicity, the r, c, d coordinate is ignored in Eq. 2. For RPN and single-stage detector, all the anchor points are regularized for consistency; even some of them may not be assigned labels according to the simple IOU threshold matching strategy. In the second-stage detector framework, the proposals are first filtered by NMS and Top-K selection, which is also a default operation in the supervised branch [13, 24] (typically 1000 proposals left for Faster-RCNN FPN). Then the coordinates of the proposals predicted on the resized image are scaled up by 2^s times to match the original image, and vice versa. The proposals from the image pair are simply concatenated as a new proposal set for refining bounding boxes and predicting classification scores. For the second stage of Faster-RCNN, all the predictions of the proposal pairs are regularized by scale consistency loss in a similar way as shown in Eq. 2. It is worth noting that, in implementing a two-stage detector, the RoI-Pooling operator may extract features from the same level for the proposal pair, which is slightly different from single-stage detectors. Nevertheless, this operation shares the same core idea that the detector is supposed to be scale consistent.

3.2. Self-Distillation

Knowledge distillation improves generalization by replacing hard label supervision with soft label predicted by a stronger teacher model. Based on the observation, the teacher model is implemented as an exponential moving average (EMA) of the detector, which is shown to produce a model with better generalization than the student model [22, 33]. The input of the teacher model is weakly augmented. Furthermore, the model is supposed to predict consistently for similar data points. The student model is input with the strongly augmented image to propagate label to neighbor points in the semantic manifold space. For simplicity, the strong augmentation is only composed of color transformation and Cutout [6], which doesn't contain the geometric transformation. The self-distillation loss is formulated as

$$\begin{aligned} L^i_{\text{distll}} =& D_{\text{KL}}(\text{sg}(P^i(X', \theta_t)), P^i(X, \theta_s)) \\ &+ ||\text{sg}(R^i(X', \theta_t)) - R^i(X, \theta_s)||_2, \end{aligned} \tag{3}$$

where i is the i-th anchor box, X' is the weakly augmented image and X is the strongly augmented image. P and R represent the classification score and bounding box regression same as in Eq. 2. The slowly progressing teacher model weights θ_t are updated from the student model weights θ_s every iteration,

$$\theta_t = \alpha\theta_t + (1 - \alpha)\theta_s. \tag{4}$$

Self-Distillation loss constrains each anchor point for RPN and one-stage detector, similar to Scale Consistency Regularization. In the scenario of the two-stage detector, all the proposals are simply concatenated as a new proposal set. Similar to Scale Consistency Regularization, all the predictions of RoIs are regularized as Eq. 3.

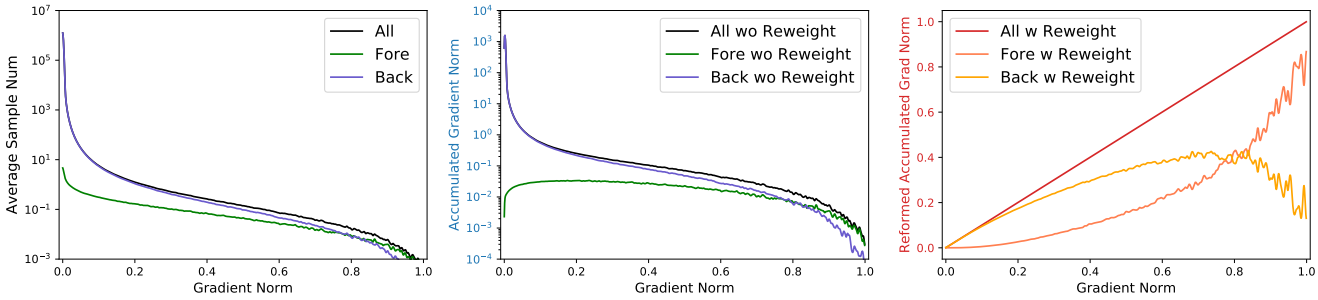


Figure 4. The average sample is the average anchor number in a single image. The vanilla method is simply treating all samples equally. The samples with large gradients do not contribute significantly because the sample number is relatively small. Our re-weighting strategy focuses on the samples with large score discrepancies and linearizes the relationship between gradient contribution and score distance.

3.3. Re-weighting Strategy

One-stage object detection methods, like RetinaNet [17] and RPN [24], face an extremely class imbalance during training. Due to the overwhelming background samples, most objectness scores are close to 0. Therefore, the KL divergence between the target and source distribution in Eq. 2 and Eq. 3 is close to 0 for most anchor boxes. Simply averaging the unsupervised loss leads to the easy samples contributing significantly to the gradient, as illustrated in Fig. 4. We aim to reduce the discrepancy between similar unlabeled inputs, especially for the potential foreground instances predicted with high objectness scores. In other words, the hard examples should contribute to the gradient more than the easy examples. Inspired by the Gradient Harmonizing Mechanism [15], we re-weight the KL-divergence by the sample numbers in a gradient range to build a linear relationship between the gradient norm and the integral gradient contribution, as illustrated in Fig. 4. Specifically, the gradient of the logits z with KL divergence loss between probability vector p and target probability vector p' is $g = \sum_{i=1}^{C} |p_i - p'_i|$, where C is the length of probability vector. Then a histogram is constructed by splitting the gradient range $[0, 1]$ into M bins equally. The number of samples in the j-th bin is denoted as R_j, and the index of the bin where gradient g is located is defined as $idx(g)$. Finally, we have the loss function on N samples:

$$L = \frac{1}{M} \sum_{i=1}^{N} \frac{D_{\mathrm{KL}}(p'_i, p_i)}{R_{\mathrm{idx}(g_i)}}. \tag{5}$$

As the main bottleneck is detecting objects from the background rather than regression, only the classification loss is re-weighted by the above strategy in scale consistency loss and self-distillation loss. Our goal is to enlarge the contribution from the samples with significant discrepancies. The other methods to solve the class imbalance problem may also improve the performance.

Method	Data	LR	Iter	AP_{50}
Supervised	VOC07	0.01	40k	74.30
STAC [30]	VOC07+12	0.001	180k	77.45
DGML [34]	VOC07+12	-	-	78.60
UBT [19]	VOC07+12	0.01	180k	77.37
ISMT [36]	VOC07+12	-	-	77.23
IT [39]	VOC07+12	0.01	180k	78.30
Ours	VOC07+12	0.01	**40k**	**80.60**

Table 1. Results on Pascal VOC 2007 test set. For all the semi-supervised methods, Pascal VOC 2012 train set is treated as unlabeled data. Iter means the total training iterations. "-" means that the results or training details are missing in the source paper.

4. Experiments

Datasets. We mainly verify the validity of our method on the challenging objective detection dataset MS-COCO [18], which contains 80 object categories with about 118k images for training and 5k images for validation. For a fair comparison, we follow the experimental setup as in the previous works [19, 30, 32, 34, 39]. In particular, there are three experimental settings: (1) *PASCAL VOC*: the VOC07 [7] *trainval* set is used as the labeled dataset and the VOC12 *trainval* set is used as the unlabeled dataset , as described in Sec.3. The performance is evaluated on the VOC07 test set. VOC07 *trainval* and VOC12 *trainval* contains 5,011 and 11,540 images respectively, resulting in a roughly 1:2 ratio of labeled data to unlabeled data. (2) *COCO-standard*: we randomly sample 5 and 10% of MS-COCO 2017 training data as the labeled dataset and treat the rest of the training data as the unlabeled dataset. As the COCO train dataset contains 118k images and is class-imbalanced, some categories are composed of less than 500 instances. When data percent is 0.5% and 1%, there are only less than 5 instances in the labeled dataset for these categories. This setting is

Method	Data Percent			LR	Iteration	Stages
	5%	10%	100%			
Supervised	18.47	23.86	38.40	0.02	180k	-
STAC [30]	24.38(+5.91)	28.64(+4.78)	-	0.01	180k	Two
Unbiased Teacher [19]	27.84(+9.37)	31.39(+7.53)	-	0.01	180k	Single
Instant Teacher [39]	26.75(+8.28)	30.40(+6.54)	40.20(+1.80)	0.01	180k	Single
Interactive Teacher [36]	26.37(+7.90)	30.53(+6.67)	39.64 (+1.24)	-	-	Single
Multi-Phase Learning [34]	-	-	40.30 (+1.90)	-	-	Three
Ours	**29.01(+10.54)**	**34.02(+10.16)**	**41.50(+3.10)**	0.01	180k	Single
Supervised	-	-	40.20	0.02	270k	-
STAC [30]	-	-	39.21(-0.99)	0.01	540k	Two
Unbiased Teacher [19]	-	-	41.30(+1.10)	0.01	270k	Single
Ours	-	-	**43.40(+3.20)**	0.02	270k	Single

Table 2. Results on MS-COCO 2017 val set. For 5% and 10% protocols, the results are the mean over 5 data folds. Stages are the number of training phases. For example, STAC has two stages: train a teacher model first to hard pseudo-label and train a student model with both labeled and pseudo-labeled data. "-" means that the results or training details are missing in the source paper.

more like few-shot learning than semi-supervised learning. Therefore, we do not report the performance. For the 100% data training setting, the whole training set is used as the labeled dataset, and the additional 123k unlabeled images are used as the unlabeled dataset. The model is tested on the MS-COCO 2017 validation set. (3) *COCO-35k*: we use the 35k subset of MS-COCO 2014 validation set as the labeled dataset and the 80k training set as the unlabeled dataset. The result is reported on the MS-COCO 2014 minival set.

Implementation Details. Following STAC [30], we use Faster-RCNN [24] with FPN [16] and ResNet-50 backbone as our default object detector. The weights of the backbone are initialized by the corresponding ImageNet-Pretrained model, which is a default setting in the existing works [14, 19, 30, 39]. The stem and first stage of the backbone are frozen, and all BatchNorm layers are in *eval* mode. The weak data augmentation only contains random resize from (1333, 640) to (1333, 800) and random horizontal flip. The strong data augmentation comprises random Color Jittering, Grayscale, Gaussian Blur, and Cutout [6], without any geometric augmentation. More training and data augmentation details are in the Appendix.

4.1. Results

Pascal VOC. In Tab. 1, our method outperforms both previous multi-stage methods and single-stage methods by a large margin. Our model achieves 80.6% AP with 6.3% gain from additional VOC2012 data. In the meantime, our proposed method requires fewer training iterations, showing that our approach is effective yet efficient. Besides, our augmentation only contains color transformation without any geometric transformation or strong regularization, such as Mixup [37] and DropBlock [9].

COCO-standard. Given the whole training set, our method even further improves the strong baseline by 3.2 mAP. For a fair comparison, the learning rate and training iterations are listed in Tab. 2. Our method surpasses the previous methods under different settings of the ratio of labeled data to unlabeled data, from roughly 1:1 to 1:20, on the class-imbalanced MS-COCO dataset. Note that UBT uses Focal Loss to handle the class imbalance issue among ground truths, while we adopt the original Faster-RCNN implementation, standard cross-entropy loss. Our method focuses on the imbalance problem between foreground and background, which is more general in practice. Especially, SED achieves more than 10 mAP improvements against the supervised baseline when using 5% and 10% labeled MS-COCO data. With 10% labeled data, the performance of SED is comparable to the fully supervised baseline model.

Method	SUP	DD [23]	DGML [34]	Oracle	Ours
mAP	31.3	33.1	35.2	37.4	**38.1**

Table 3. Results on MS-COCO 2014 minival set. SUP is to train the model only with the labeled data. Oracle means treating all the 115k images as labeled data and training with only the supervised loss.

COCO-35k. MS-COCO 2014 minival set is identical to MS-COCO 2017 val set. Tab. 3 shows that our method even outperforms the Oracle result with only 35k labeled data, benefiting from the scale consistency regularization, self-distillation, and strong augmentation. The promising result indicates that the semi-supervised method can achieve a comparable result to a fully supervised counterpart.

4.2. Ablation Study

Scale Consistency Regularization constrains the discrepancy between the predictions of images of different

Method	SCR	Self-Distill		Reweight	mAP
		Hard	Soft		
SUP					23.86
Ours	✔				26.80
	✔			✔	30.10
			✔		29.80
			✔	✔	31.40
	✔	✔		✔	29.50
	✔		✔	✔	34.00

Table 4. The ablative results on MS-COCO 2017 val set. The models are trained with 10% labeled and 90% unlabeled MS-COCO train 2017 split. The SCR represents the scale consistency regularization. We test the self-distillation with two types of target: hard target and soft target.

sizes. By comparing the second row in Tab. 4 with baseline, we find that Scale Consistency Regularization improves about 3 mAP without our re-weighting strategy, naively averaging the loss across the anchor boxes and RoIs. Although suffering from the class imbalance problem, Scale Consistency Regularization is promising. Fig. 2b shows that the discordance between different sizes is alleviated.

Self-Distillation with Soft Target surpasses the hard pseudo-label counterpart over 4.5 mAP, which demonstrates that the quality of hard pseudo-label is inferior. Self-Distillation gains about 6 mAP against the baseline individually. The soft target method benefits from fewer False Negative samples and the structural information via knowledge distillation. Furthermore, our approach based on soft target is threshold-free, which is simpler and easier to transfer to other datasets.

Re-weighting Strategy focuses on the anchor or RoI pairs with large discrepancies and transforms the relationship between gradient contribution and score distance to linearity. The results of Scale Consistency Regularization and Self-Distillation with Soft Target are increased by 3.3 mAP and 1.6 mAP separately. For Faster-RCNN, our re-weighting strategy still takes effect even though the RoIs are predicted after NMS and Top-K selection operation, increasing the foreground to background sample ratio.

4.3. Discussion

How to Extend SED to Other Detectors. Most detectors (e.g. RetinaNet, Faster-RCNN) assign foreground labels to the "anchor box" according to a similar rule, the Intersection-over-Union (IoU) threshold criterion. For DETR [3], a single feature map detector, we match the predictions of input in different views according to Hungarian algorithm, where the pair-wise matching cost is defined as: $L_{\text{match}} = D_{\text{JS}}(p_1, p_2) + \lambda L_{\text{IoU}}(b_1, b_2)$, where $D_{\text{JS}}(p_1, p_2)$ is JS-Divergence between the probability vectors and L_{IoU} is GIoU loss [25]. According to the above analysis, our method can be extended to RetinaNet and DETR with different backbones. The results in Tab. 5 demonstrate that our method is valid for different classes of the detector.

Model	Retina w R50	Retina w R18	DETR w R50
SUP	23.6	21.5	64.9
Ours	33.0	31.4	69.3

Table 5. For RetinaNet, the experiments are conducted on MS-COCO set with 10% labeled training data. Due to the extremely long training epoch of DETR, we report the result on Pascal VOC 2007 test set. Both supervised and our DETR are trained for 300 epochs.

Range	[640, 800]	[300, 1200]	Ours [640, 800]
Result	31.4	32.0	34.0

Table 6. The scale jittering results on MS-COCO 2017 val set. The models are trained with 10% labeled and 90% unlabeled MS-COCO train 2017 split. Range is the range of the short edge. The results show that the scale consistency loss is beyond large scale jittering augmentation.

Relationship with Large Scale Jittering. The proposed scale consistency regularization is more than large scale jittering augmentation. The object of our method is $L = L(x) + L(x') + L_{\text{scr}}(x, x')$, while the object of a large scale jittering augmentation is $L = L(x) + L(x')$, where x and x' are the input image in different views. The L_{scr} is the scale consistency loss. The constraint of our method is stronger than large scale jittering augmentation. Thus we believe that the parameter space of local minimum is a subset of that of large scale jittering. Tab. 6 also shows that our method encourages the model to converge with less generalization error.

Relationship with Multi-Scale Testing. Tab. 7 shows that the baseline models benefit from multi-scale testing by an ensemble with NMS (Threshold=0.5). The model trained with 10% labeled data is increased by 2.0 mAP, and the fully supervised model (SUP 100%) gets 1.5 mAP improvement. However, this improvement comes from the discrepancy between the predictions of images in different sizes. Moreover, the ensemble method also consumes $2.5\times$ more inference time than the single-scale testing method. Our method benefits less from multi-scale testing as a consequence of the proposed scale consistency regularization. Our method significantly improves the single-scale testing performance, which has more practical value.

Downsampling Rate in Scale Consistency Regularization. As shown in Tab. 8, the model achieves the best result when the downsampling rate is set to 2 (i.e. the S in

Model	Image Size			Ensemble
	480	800	1200	
SUP 10%	22.9	24.1	22.5	26.1(+2.0)
SUP 100%	33.7	37.4	36.8	38.9(+1.5)
Ours 10%	31.5	34.2	33.0	34.8(+0.6)

Table 7. Multi-Scale Testing on MS-COCO 2017 val set. The small gain indicates that the detector consistently predicts images in different sizes, which means robust to scale variance.

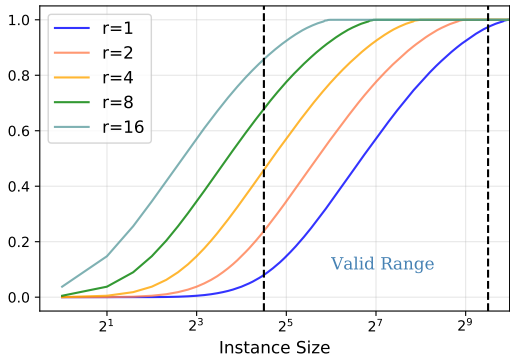


Figure 5. The CDF of instance size for Faster-RCNN detector on MS-COCO train dataset. The range of valid instance is calculated according to the label assignment rule.

Sec. 3.1 is set to 1.). The performance is inferior as the downsampling rate scales up, which means that regularizing the scale consistency with too small images is less effective. The anchor-based detector refines the prior boxes, which constrains the valid detection scale range (from 22.6 to 724.1 pixel^2, theoretically). Fig. 5 shows that the fraction of instances in the valid range is highest when the downsampling rate is set to 2. All the models are trained with 10% COCO training data, using RetinaNet with FPN and ResNet-18 backbone in Tab. 8.

Exponential Moving Average (EMA) Rate of Teacher model. In Eq. 4, the weight of the teacher is updated in an exponential moving average manner. The EMA update can be viewed as the average weight of the models in the past $\frac{\alpha}{1-\alpha}$ steps approximately. As the learning rate policy is *step*, which decays the learning rate by 0.1 at each milestone iteration, the performance of the teacher is inferior to the student model after switching the learning rate, which leads to the degradation of the student model. We observe the same appearance in UBT [19], which sets the α to 0.9996 and adopts *step* learning rate policy. To alleviate the degradation, we propose to decay the EMA update rate at the same milestone iteration as the learning rate. The results in Tab. 9 show that our *step* decay method and *cosine* decay method both surpass the baseline model.

Compare Re-weighting strategy with other methods. We conduct experiments by replacing our re-weighting strategy with OHEM (Online Hard Example Mining) and Focal Loss [17]. The vanilla method is training without any class balancing technique. The results in Tab. 10 show that our method is effective.

Rate	mAP
1	23.0
2	**26.1**
4	25.2
8	23.1
16	21.1

Table 8. Results on COCO val set. Rate is the downsampling rate.

Start	End	Policy	mAP
0.996	0.9	Cosine	33.0
0.99	0.9	Step	**34.1**
0.95	0.95	None	32.0

Table 9. Results on COCO val set. Start and End mean the initial EMA update rate and the target rate. Cosine policy is cosine annealing schedule. Our Step policy only decays once at the first milestone iteration.

Method	Vanilla	OHEM [32]	Focal [17]	Ours
Result	30.1	31.4	31.2	34.0

Table 10. The comparison results of re-weight strategy on MS-COCO 2017 val set. The Faster-RCNN models are trained with 10% labeled and 90% unlabeled MS-COCO train 2017 split.

5. Conclusion

In this work, we introduce a novel semi-supervised object detection framework based on the consistency regularization method. Our scale consistency regularization overcomes the large scale variance challenge and significantly improves the performance on single-scale testing. Further, SED alleviates the negative effect of False Negative samples and benefits from the structural information via knowledge distillation. The re-weighting strategy focuses on the potential fore-ground regions of the unlabeled data and linearizes the relationship gradient contribution and score distance. Experiments on MS-COCO and Pascal VOC show that Scale-Equivalent Distillation significantly improves the performance with different ratios of labeled data to unlabeled data and can be extended to different detector classes. Our framework is a holistic approach compatible with other semi-supervised methods, such as Mixmatch and Noisy student self-distillation. In addition, Scale-Equivalent Distillation framework could be further extended to other dense prediction tasks, like instance segmentation, joint human parsing, and post estimation. Our method has great potential to promote the development of semi-supervised learning and further reduce the dependence of labeled data with no negative social impact.

Acknowledgement. Ping Luo is supported by the General Research Fund of HK No.27208720 and 17212120.

References

[1] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. 2020. 2

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 2

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 7

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020. 3

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 3

[6] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 4, 6

[7] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 2, 5

[8] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *International Conference on Machine Learning*, pages 1607–1616. PMLR, 2018. 3

[9] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Dropblock: A regularization method for convolutional networks. *arXiv preprint arXiv:1810.12890*, 2018. 6

[10] Qiushan Guo, Xinjiang Wang, Yichao Wu, Zhipeng Yu, Ding Liang, Xiaolin Hu, and Ping Luo. Online knowledge distillation via collaborative learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11020–11029, 2020. 3

[11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 3

[12] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4918–4927, 2019. 3

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 4

[14] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. *Advances in neural information processing systems*, 32:10759–10768, 2019. 2, 6

[15] Buyu Li, Yu Liu, and Xiaogang Wang. Gradient harmonized single-stage detector. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8577–8584, 2019. 5

[16] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 4, 6

[17] 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*, pages 2980–2988, 2017. 2, 3, 5, 8

[18] 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*, pages 740–755. Springer, 2014. 1, 2, 5

[19] 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. *arXiv preprint arXiv:2102.09480*, 2021. 2, 5, 6, 8

[20] 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*, 41(8):1979–1993, 2018. 2

[21] Avital Oliver, Augustus Odena, Colin Raffel, Ekin D Cubuk, and Ian J Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. *arXiv preprint arXiv:1804.09170*, 2018. 2

[22] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992. 4

[23] Ilija Radosavovic, Piotr Dollár, Ross Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omni-supervised learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4119–4128, 2018. 6

[24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015. 3, 4, 5, 6

[25] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 658–666, 2019. 7

[26] Laine Samuli and Aila Timo. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations (ICLR)*, volume 4, page 6, 2017. 2

[27] Bharat Singh and Larry S Davis. An analysis of scale invariance in object detection snip. In *Proceedings of the*

IEEE conference on computer vision and pattern recognition, pages 3578–3587, 2018. 2

[28] Bharat Singh, Mahyar Najibi, and Larry S Davis. Sniper: Efficient multi-scale training. *arXiv preprint arXiv:1805.09300*, 2018. 2

[29] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020. 1, 2

[30] 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, 6

[31] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017. 3

[32] 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*, pages 3132–3141, 2021. 2, 5, 8

[33] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1195–1204, 2017. 1, 2, 3, 4

[34] Zhenyu Wang, Yali Li, Ye Guo, Lu Fang, and Shengjin Wang. Data-uncertainty guided multi-phase learning for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4568–4577, 2021. 5, 6

[35] 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*, pages 10687–10698, 2020. 1

[36] 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*, pages 5941–5950, 2021. 2, 5, 6

[37] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 6

[38] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4320–4328, 2018. 3

[39] 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*, pages 4081–4090, 2021. 2, 5, 6

[40] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin D Cubuk, and Quoc V Le. Rethinking pre-training and self-training. *arXiv preprint arXiv:2006.06882*, 2020. 3