SCALE-INVARIANT TEACHING FOR SEMI-SUPERVISED OBJECT DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

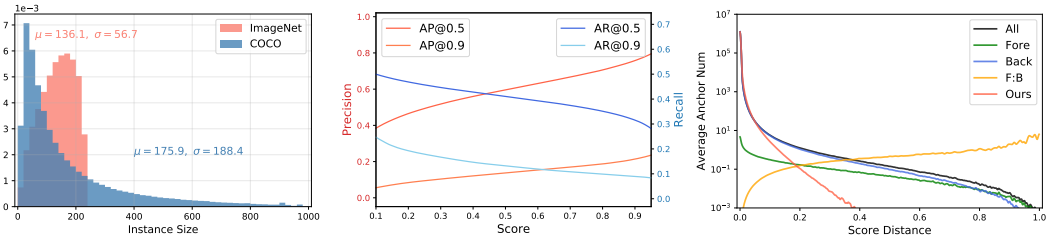
Recent Semi-Supervised Object Detection methods are mainly based on self-training, i.e., generating hard pseudo-labels by a teacher model on unlabeled data as supervisory signals. Although they achieved certain success, the massive False Negative samples and inferior localization precision lack consideration. Furthermore, the limited annotations in semi-supervised learning scale up the challenges: large variance of object sizes and class imbalance (i.e., the extreme ratio between background and object), hindering the performance of prior arts. We address these challenges by introducing a novel approach, Scale-Invariant Teaching (SIT), which is a simple yet effective end-to-end knowledge distillation framework robust to large object size variance and class imbalance. SIT has several appealing benefits compared to previous works. (1) SIT imposes a consistency regularization to reduce the prediction discrepancy between objects with different sizes. (2) The soft pseudo-label 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 from unlabeled data to reduce the effect of class imbalance. Extensive experiments show that SIT consistently outperforms the recent state-of-the-art methods and baseline 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 (about 1.28 million labeled images). However, different from classification, object detection further involves locating the presence of objects with a bounding box. Therefore, the annotation for object detection is much more expensive, leading to labeled data remaining scarce related to classification. As an alternative, Semi-supervised Learning (SSL) improves a model's performance significantly by leveraging both the limited labeled data and the massive unlabeled data.

Recently, Semi-Supervised Learning for classification has received much attention (Tarvainen & Valpola, 2017; Berthelot et al., 2019; Xie et al., 2020; Sohn et al., 2020a). However, Semi-Supervised Object detection (SS-OD) is more challenging than Semi-Supervised Image Classification. The reason is threefold. The scale of objects varies in a small range for classification, whereas the scale variation is large across object instances in MS-COCO dataset (Lin et al., 2014). As shown in Fig. 1a, the standard variance of the scale of instances in MS-COCO is 188.4, while that of ImageNet is 56.7. Besides, the hard pseudo-label predicted by Self-training methods incurs noise because of inaccurate bounding box regression and False Negative object instance. As illustrated in Fig. 1b, 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. Yet, the foreground and background classes are high imbalanced in object detection. The ratio of the foreground sample to that of the background sample is about $1:25,000$ under RetinaNet framework (Lin et al., 2017b).

A detector is supposed to be scale-invariant to object instances, which means that the predictions of an image in different sizes should be consistent. However, we observe a discrepancy in the objectness scores, as indicated in Fig. 1c. The ratio of foreground anchor to background anchor



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

Figure 1: (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. (b) 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. (c) All the scores are predicted by the RetinaNet detector with FPN and ResNet 50 backbone, which is trained with randomly sampled 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.

increases as the score distance becomes large, which implies that the network detects an object instance in the image while is blind to the instance in a different size, in the case of a significant objectness score distance. This inconsistency is typically alleviated by the multi-scale inference ensemble, which increases the computational cost and requires complicated operations to fuse the bounding boxes.

STAC (Sohn et al., 2020b) simply adopts the existing advanced semi-supervised image classification method to solve Semi-Supervised Object Detection straightly, as illustrated in Fig. 2a. UBT (Liu et al., 2021) adopts Focal Loss (Lin et al., 2017b) to fix the overfitting hard pseudo-label issue, as shown in Fig. 2b. Whereas the performance of bot STAC and UBT is moderate in the high-data scenario due to the False Negative object instance and inferior localization precision.

To overcome the challenges motioned above, we propose Scale-Invariant Teaching, a simple yet effective end-to-end semi-supervised learning framework. Since the scale is an essential dimension of the low-dimensional semantic manifolds, we design a scale-invariant consistency regularization across feature maps 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 slowly progressing teacher is proposed to generate soft pseudo-labels for unlabeled data in an online manner. The teacher is implemented as an exponential moving average (EMA) of the detector, which doesn't increase the learnable parameters. Weight averaging is shown to improve generalization performance (Tarvainen & Valpola, 2017; Athiwaratkun et al., 2018), yielding a stronger teacher than the student model. Considering the class imbalance problem, we implement a re-weighting strategy to focus on the inconsistency among feature maps in different levels and the discordance between EMA 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 effect of SIT, we conduct extensive experiments on benchmarks for object detection, MS-COCO (Lin et al., 2014) and Pascal VOC (Everingham et al., 2010). We demonstrate that our method surpasses the baseline model and previous methods by large margins, even outperforming the fully supervised counterpart with 35k MS-COCO labeled data.

Our contributions are listed as follows: (1) SIT imposes a scale-invariant consistency regularization to reduce the prediction discrepancy between objects with different sizes. (2) The soft pseudo-label 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-Learning. 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 classification tasks, Self-training methods (Tarvainen & Valpola, 2017; Berthelot et al., 2019; 2020; Sohn et al., 2020a) performs well. However, Semi-Supervised Object detection is more challenging than Semi-Supervised Image Classification.

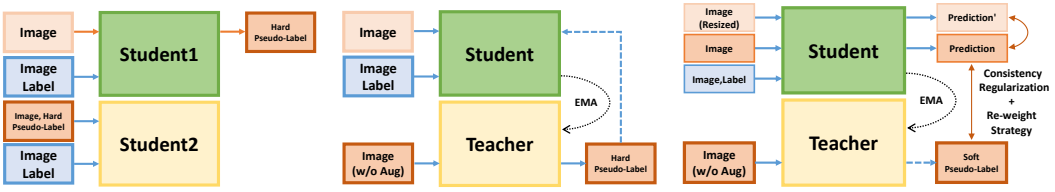


Figure 2: The Image (w/o Aug) means that the input is only weakly augmented (random resize and flip). The Image in (c) is strongly augmented, such as color jittering, gaussian blur. (a) In STAC (Sohn et al., 2020b), a model is trained with labeled data to predict hard pseudo-label for self-training during the first stage. (b) Unbiased Teacher (Liu et al., 2021) generates hard pseudo-label by the slowly progressing teacher, which is shown to yield more accurate targets (Tarvainen & Valpola, 2017). (c) Our model improves the scale invariance, which is critical for object detectors, by regularizing the consistency between different-sized images. Furthermore, the inherent False Negative sample noise is alleviated by predicting soft pseudo-label. A re-weighting strategy is adopted to solve the severe class imbalance problem.

Some works (Liu et al., 2021; Zhou et al., 2021) focused on SS-OD 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 on the foreground and background classification and refine the hard pseudo-labels by ensemble methods. Nevertheless, methods based on hard pseudo-label have an inherent defect that False Negative object instances, especially those whose scores are near the threshold, influence the consistency of recognition. Humble Teacher (Tang et al., 2021) 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 magnitude of gradients from the two kinds of regions is quite different. Therefore, the regions should be treated with different importance. Different from the existing works, our method generates soft pseudo-labels for unlabeled data in an online manner, and the re-weighting strategy automatically focuses on the potential foreground regions from 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 translations of the same image should result in similar predicted probabilities. This class of methods (Samuli & Timo, 2017; Tarvainen & Valpola, 2017; Miyato et al., 2018) doesn't generate pseudo-label but constrains the discrepancy between the outputs, which is known to help smooth the manifold (Oliver et al., 2018). For SS-OD, CSD (Jeong et al., 2019) 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 (Lin et al., 2014) detection dataset, the scale of the smallest and largest 10% of object instances is 0.024 and 0.472, respectively, which results in scale variations of almost 20 times. Our method regularizes the feature map in different sizes to solve the large scale variation. Furthermore, the consistency regularization with EMA teacher (Tarvainen & Valpola, 2017) is self-distillation (Furlanello et al., 2018; Zhang et al., 2018; Guo et al., 2020) from the perspective of soft targets, which benefit from high-quality prediction.

Pre-Training. In recent years, it has been a paradigm that pre-train backbone on a large-scale dataset, such as ImageNet (Deng et al., 2009) or JFT (Sun et al., 2017), 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 scenario of small data (He et al., 2019; Zoph et al., 2020), which is an extreme of semi-supervised learning. SimCLR (Chen et al., 2020) and MOCO (He et al., 2020) 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-Invariant Teaching

Problem Definition. Semi-supervised learning is halfway between supervised and unsupervised learning. More precisely, our model is trained with a labeled set $D_s = \{x_i^s, y_i^s\}_{i=1}^{N_s}$ and an unlabeled

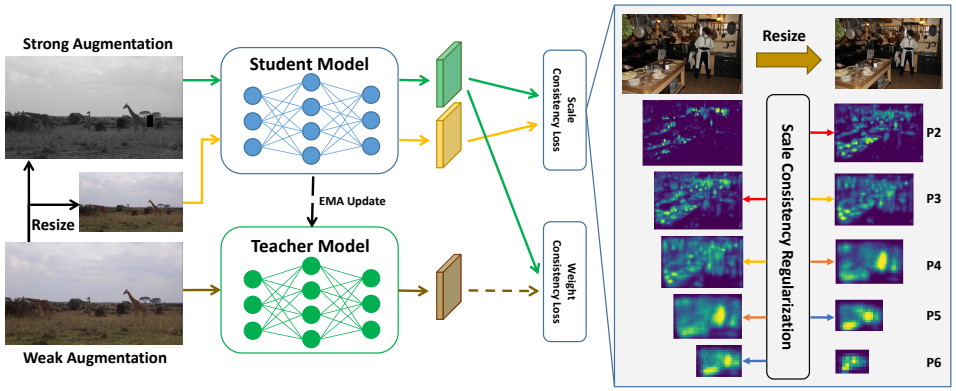


Figure 3: Overview of our method. For simplicity, the supervised branch is ignored, which shares the Student Model with the unsupervised branch. The dashed line means the prediction of the Teacher Model is not optimized by the gradient. For Scale consistency Regularization, the loss constrains predictions from different levels, linked by arrows of the same color (best viewed in color).

set $D_u = \{x_i^s\}_{i=1}^{N_u}$, 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-Invariant Teaching consists of two branches, the supervised branch and the unsupervised branch, as illustrated in Fig. 3. The supervised branch is trained by following the normal procedure, like (Ren et al., 2015; Lin et al., 2017b). The unsupervised branch is under a teacher-student framework, in which the teacher is implemented as an exponential moving average of the student. SIT aims to predict consistently for the scale variants of input. In practice, the student processes the strongly augmented unlabeled images and resized weakly augmented 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. As the teacher is updated from the student weights, the constraint is viewed as a weight consistency regularization for aligning the name of the unsupervised loss. The final loss is the weighted sum of the supervised loss and unsupervised loss,

$$L = L_s + \frac{n_u}{n_s}(\lambda_s L_{scale} + \lambda_w L_{weight}), \tag{1}$$

where n_u, n_s are the batch size of unlabeled data and labeled data, L_{scale} and L_{weight} are Scale-Invariant Consistency Regularization and Weight Consistency Regularization. For two-stage detectors, the unsupervised losses are applied to both RPN head and ROI head.

3.1 Scale-Invariant Consistency Regularization

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

As indicated in Fig. 3, Scale Consistency Loss regularizes feature maps 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 image is downsampled to $\frac{1}{2^s}$ original size. Towards handling the large scale variation, the s is selected from $\{1, 2, 3, 4\}$, 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 equivalently for the corresponding levels.

Precisely, the Scale-Invariant Consistency Loss is defined as

$$L^f_{scale} = KL(sg(P^f(X)), P^{f'}(\hat{X}')) + KL(sg(P^{f'}(\hat{X}')), P^f(X)) + ||R^f(X) - R^{f'}(\hat{X}')||_2, \quad (2)$$

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 (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 the refined bounding boxes and classification scores. All the proposal pairs are regularized by Scale-Invariant 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. But this operation shares the same core idea that the detector is supposed to be scale-invariant.

3.2 Weight Consistency Regularization

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 uses the EMA weights of the student model, which is shown to produce a model with better generalization than the student model (Polyak & Juditsky, 1992; Tarvainen & Valpola, 2017). To ensure the quality of the soft pseudo-label, 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 (DeVries & Taylor, 2017), which doesn't torture the geometric information. The weight consistency loss is formulated as

$$L^i_{weight} = KL(sg(P^i(X, W_t)), P^i(X', W_s)) + ||sg(R^i(X, W_t)) - R^i(X', W_s)||_2, \quad (3)$$

where i is the i-th anchor box, X and X' is the weakly augmented image and 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 W_t are updated from the student model weights W_s every iteration,

$$W_t = \alpha W_t + (1 - \alpha) W_s. \quad (4)$$

Similar to Scale-Invariant Consistency Regularization, Weight Consistency Regularization is applied to each anchor point for RPN and one-stage detector. In the scenario of the two-stage detector, all the proposals passed NMS and Top-K selection are simply concatenated as a new proposal set. All the predictions of RoIs are regularized as Eq. 3.

3.3 Re-weighting Strategy

One-stage object detection methods, like RetinaNet (Lin et al., 2017b) and RPN (Ren et al., 2015), 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 target distribution and source distribution in Eq. 2 and Eq. 3 is close to 0 for most anchor boxes. Simply averaging the Consistency 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 (Li et al., 2019), 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 KL divergence 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

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. AP_{50} is reported. LR is the learning rate, and Iter means the total training iterations. C indicates the Color transformation augmentation, G is the Geometric transformation augmentation, and Mosaic is randomly performing horizontal mixing and vertical mixing two images. Mixup (Zhang et al., 2017) and DropBlock (Ghiasi et al., 2018) are strong regularization operations.

Method	Data	LR	Iter	AP	Augmentation
Supervised	VOC07	0.01	40k	74.3	-
STAC (Sohn et al., 2020b)	VOC07+12	0.001	180k	77.45	C, G
DGML (Wang et al., 2021)	VOC07+12	-	-	78.60	-
UBT (Liu et al., 2021)	VOC07+12	0.01	180k	77.37	C
ISMT (Yang et al., 2021)	VOC07+12	-	-	77.23	C, DropBlock
IT (Zhou et al., 2021)	VOC07+12	0.01	180k	78.30	C, Mixup, Mosaic
Ours	VOC07+12	0.01	**40k**	**80.60**	**C**

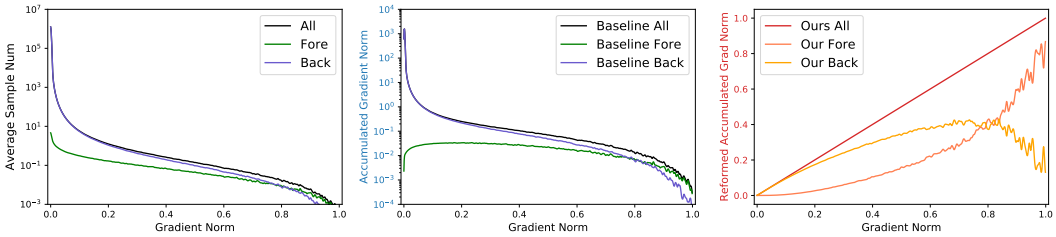


Figure 4: The average sample is the average anchor number in a single image. The baseline method is simply treating all samples equally. The samples with large gradients don't contribute significantly because the sample number is relatively small. Our re-weighting strategy focuses on the samples with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

gradient g is located is defined as $idx(g)$. Finally, we have the loss function:

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

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-Invariant Consistency Loss and Weight Consistency 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.

4 Experiments

Datasets. We mainly verify the validity of our method on the challenging objective detection dataset MS-COCO (Lin et al., 2014), 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 the previous work (Sohn et al., 2020b; Liu et al., 2021; Zhou et al., 2021; Tang et al., 2021; Wang et al., 2021). In particular, there are three experimental settings: (1) *PASCAL VOC*: the VOC07 (Everingham et al., 2010) *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. Besides, the whole training set is used as the labeled dataset, and the additional 123k unlabeled images are used as the unlabeled dataset, which is denoted as 100% data training setting. 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 performance is reported on the MS-COCO 2014 minival set (5000 images).

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 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 original paper. For 180k training schedule, the learning rate is set to 0.01 with 5% and 10% data protocol, to 0.02 with 100% data protocol.

Method	Data Percent			LR	Iter	Stages
	5%	10%	100%			
SUP	18.47	23.86	38.40	0.02	180k	-
STAC	24.38(+5.91)	28.64(+4.78)	-	0.01	180k	Two
UBT	27.84(+9.37)	31.39(+7.53)	-	0.01	180k	Single
IT	26.75(+8.28)	30.40(+6.54)	40.20(+1.80)	0.01	180k	Single
ISMT	26.37(+7.90)	30.53(+6.67)	39.64	-	-	Single
DGML	-	-	40.30	-	-	Three
Ours	**29.01(+10.54)**	**34.02(+10.16)**	**41.50(+3.10)**	0.01 / 0.02	180k	Single
SUP	-	-	40.20	0.02	270k	-
STAC	-	-	39.21(-0.99)	0.01	540k	Two
UBT	-	-	41.30(+1.10)	0.01	360k	Single
Ours	-	-	**43.40(+3.20)**	0.02	270k	Single

Table 3: Results on MS-COCO 2014 minival set. DD is Data Distillation (Radosavovic et al., 2018). Oracle means treating all the 115k images as labeled data.

Method	Baseline	DD	DGML	Oracle	Ours
AP	31.3	33.1	35.2	37.4	**38.1**

Implementation Details. Following STAC (Sohn et al., 2020b), we use Faster-RCNN (Ren et al., 2015) with FPN (Lin et al., 2017a) 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 existing works (Sohn et al., 2020b; Jeong et al., 2019; Liu et al., 2021; Zhou et al., 2021). The stem and first stage of the backbone are frozen, and all BatchNorm layers are in *eval* mode. For data augmentation, the weak data augmentation only contains random resize from (1333, 640) to (1333, 800) and random horizontal flip with a probability of 0.5. The strong data augmentation is composed of random Color Jittering, Grayscale, Gaussian Blur, and Cutout (DeVries & Taylor, 2017), 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 is simply applying color transformation without any geometric transformation or strong regularization, such as Mixup (Zhang et al., 2017), DropBlock (Ghiasi et al., 2018).

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 the 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, Scale-Invariant Teaching 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 Scale-Invariant Teaching is comparable to the fully supervised baseline model. This phenomenon demonstrates that the consistency-based semi-supervised learning method exploits the information of unlabeled data efficiently.

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.

Method	Scale-Consistency	Weight-Consistency		Reweight	mAP
		Hard Target	Soft Target		
SUP					23.86
Ours	✔				26.80
	✔			✔	30.10
			✔		29.80
			✔	✔	31.40
	✔	✔		✔	29.50
	✔		✔	✔	34.00

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.

4.2 Ablation Study

Scale Consistency Regularization constrains the discrepancy between the predictions of images of different sizes. By comparing the second row with baseline, we find that Scale-Consistency improves about 3 mAP without our re-weighting strategy, naively averaging the loss across the anchor boxes and RoIs. Although suffering from the foreground-background imbalance problem, Scale Consistency Regularization is promising. Fig. 1c shows that the discordance between different sizes is alleviated.

Weight Consistency Regularization with Soft Target surpasses the hard pseudo-label counterpart over 4.5 mAP, which demonstrates that the quality of hard pseudo-label is inferior. Weight Consistency Regularization 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 discrepancy and transforms the relationship between gradient contribution and score distance to linearity. The results of Scale-Consistency Regularization and Weight Consistency Regularization 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 ratio of foreground to background sample.

4.3 Discussion

Relationship with Multi-Scale Testing. The Tab. 5 shows that the baseline models benefit from multi-scale testing by a simple ensemble with NMS (Threshold=0.5). The model trained with 10% labeled data is increased by 2.0 mAP, and the fully supervised model 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 inference as a consequence of the proposed scale-invariant consistency regularization, which means the detector has strong scale invariance. Our scale-invariant consistency regularization significantly improves the single-scale testing performance, which has more practical value.

Downsampling Rate in Scale-Invariant Consistency Regularization. As shown in Tab. 6, the model achieves the best result when the downsampling rate is set to 2. The performance is inferior as the downsampling rate scales up, which means that regularizing the scale invariance with too small images is less effective. The anchor-based detector is to refine the prior bounding boxes, which constrains the valid detection scale range (from 22.6 to 724.1, theoretically). Fig. 5 shows that the fraction of valid instances 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-50 backbone.

Table 5: Multi-Scale Testing on MS-COCO 2017 val set. The ensemble results show the gain from multi-scale testing, which means the model detects instances in one size while is blind to them in the other size. The small gain indicates that the detector consistently predicts images in different sizes, which means robust scale invariance.

Model	Image Size			Ensemble
	(1333, 480)	(1333, 800)	(1400, 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)

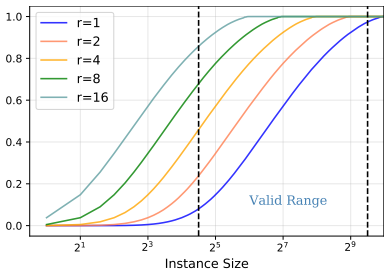
Table 6: Results on COCO val set. Rate is the down-sampling rate in Sec. 3.1.

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

Table 7: 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.

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

Figure 5: The CDF of instance size distribution on the whole MS-COCO train dataset.



EMA Rate in Weight Consistency Regularization. In Eq. 4, the weight of the teacher is the updated in an exponential moving average manner, which 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 EMA 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 (Liu et al., 2021), 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. 7 shows that our step decay method and cosine decay method both surpass the baseline model.

5 Conclusion

In this work, we introduce a novel semi-supervised object detection framework based on the consistency regularization method. Our scale-invariant consistency regularization smooths the scale manifold and significantly improves the performance on single-scale testing. Further, the weight consistency regularization benefits from the structural information via knowledge distillation and alleviates the negative effects of False Negative samples. The re-weighting strategy focuses on the sample pairs with large discrepancy of prediction and linearizes the relationship gradient contribution and score distance. Experiments on COCO and Pascal VOC show that Scale-Invariant Teaching significantly improves the performance with different ratios of labeled data to unlabeled data. Our framework is a holistic approach compatible with other semi-supervised methods, such as Mixmatch and Noisy student self-distillation. In addition, our Scale-Invariant Teaching framework could be further extended to anchor-based, anchor-free single-stage detectors and other dense prediction tasks, like instance segmentation, joint human parsing, and post estimation.

Reproducibility

As shown in the main text and the appendix, all the training details are provided to reproduce the reported results. In addition, the proposed method is simple yet efficient to implement with MMDetection framework (Chen et al., 2019). Our code will be released.

REFERENCES

Ben Athiwaratkun, Marc Finzi, Pavel Izmailov, and Andrew Gordon Wilson. There are many consistent explanations of unlabeled data: Why you should average. *arXiv preprint arXiv:1806.05594*, 2018.

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.

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.

Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

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 (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 13–18 Jul 2020. URL `http://proceedings.mlr.press/v119/chen20j.html`.

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*, pp. 248–255. Ieee, 2009.

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

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.

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

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

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*, pp. 11020–11029, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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

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*, pp. 9729–9738, 2020.

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.

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

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*, pp. 740–755. Springer, 2014.

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*, pp. 2117–2125, 2017a.

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*, pp. 2980–2988, 2017b.

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.

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.

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.

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

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*, pp. 4119–4128, 2018.

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.

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

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*, 2020a.

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*, 2020b.

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*, pp. 843–852, 2017.

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*, pp. 3132–3141, 2021.

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*, pp. 1195–1204, 2017.

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*, pp. 4568–4577, 2021.

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*, pp. 10687–10698, 2020.

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*, pp. 5941–5950, 2021.

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

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*, pp. 4320–4328, 2018.

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*, pp. 4081–4090, 2021.

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.

A APPENDIX

A.1 IMPLEMENTATION AND TRANING DETAILS.

Our implementation is based on MMDetection framework (Chen et al., 2019). The default detector is set as Faster-RCNN (Ren et al., 2015) with FPN (Lin et al., 2017a) and ResNet-50 (He et al., 2016) for a fair comparison with prior works (Sohn et al., 2020b; Yang et al., 2021; Zhou et al., 2021; Liu et al., 2021; Wang et al., 2021).

Training Details. The weights of the backbone are first initialized by the corresponding ImageNet-Pretrained model, which is a default setting in existing works (Sohn et al., 2020b; Jeong et al., 2019; Liu et al., 2021; Zhou et al., 2021). All the models are trained with learning rate starting at 0.01 and the learning rate drops by 0.1 at the 120k and 160k iteration for 180k training schedule as default. We set the weight decay to 0.0001, batch size to 16, and the momentum is 0.9 for SGD optimizer. Like (Liu et al., 2021), we separate 5k/10k/12k/90k iterations from the whole process as the burn-in phase for 5%/10%/35k/100% data protocols. For verifying the effectiveness of our method, we simply set the λ_s and λ_w in Eq. 1 as 0.5 and 1 separately. The EMA update rate starts with 0.99 and steps to 0.9 at the 120k iteration, aligned with learning rate decay policy.

Table 8: Details of data augmentations.

Strong Augmentation			
Process	Probability	Parameters	Details
Color Jittering	0.8	brightness, contrast, saturation = 0.4, 0.4, 0.4	Brightness factor is chosen uniformly from [0.6, 1.4], Contrast factor is chosen uniformly from [0.6, 1.4], Saturation factor is chosen uniformly from [0.6, 1.4]
Grayscale	0.2	None	None
GaussianBlur	0.5	$\sigma \sim U(0.1, 2.0)$	Gaussian filter kernel size is 23
Cutout 1	0.7	scale=(0.05, 0.2), ratio=(0.3, 3.3)	Randomly selects a rectangle region in an image
Cutout 2	0.5	scale=(0.02, 0.2), ratio=(0.1, 6)	Randomly selects a rectangle region in an image
Cutout 3	0.3	scale=(0.02, 0.2), ratio=(0.05, 8)	Randomly selects a rectangle region in an image

Data Augmentation. As shown in Tab. 8, the weak data augmentation only contains random resize from (1333, 640) to (1333, 800) and random horizontal flip with a probability of 0.5. The strong data augmentation is composed of random Color Jittering, Grayscale, Gaussian Blur, and Cutout (DeVries & Taylor, 2017), without any geometric augmentation.