Dense Learning based Semi-Supervised Object Detection

Binghui Chen[1], Pengyu Li[1], Xiang Chen[1], Biao Wang[1], Lei Zhang[2], Xian-Sheng Hua[1]
[1] Alibaba Group, [2] The Hong Kong Polytechnic University
chenbinghui@bupt.cn, lipengyu007@gmail.com, xchen.cx@alibaba-inc.com, wangbiao225@foxmail.com
cslzhang@comp.polyu.edu.hk, huaxiansheng@gmail.com

Abstract

Semi-supervised object detection (SSOD) aims to facilitate the training and deployment of object detectors with the help of a large amount of unlabeled data. Though various self-training based and consistency-regularization based SSOD methods have been proposed, most of them are anchor-based detectors, ignoring the fact that in many real-world applications anchor-free detectors are more demanded. In this paper, we intend to bridge this gap and propose a DenSe Learning (DSL) based anchor-free SSOD algorithm. Specifically, we achieve this goal by introducing several novel techniques, including an Adaptive Filtering strategy for assigning multi-level and accurate dense pixel-wise pseudo-labels, an Aggregated Teacher for producing stable and precise pseudo-labels, and an uncertainty-consistency-regularization term among scales and shuffled patches for improving the generalization capability of the detector. Extensive experiments are conducted on MS-COCO and PASCAL-VOC, and the results show that our proposed DSL method records new state-of-the-art SSOD performance, surpassing existing methods by a large margin. Codes can be found at https://github.com/chenbinghui1/DSL.

1. Introduction

The recent rapid development of object detection (OD) methods [5, 17, 40] largely owes to the availability of large-scale and well-annotated datasets, such as MS-COCO benchmark [27]. With the increasing demand for more powerful and accurate detection models, the need to collect and label more data also increases. However, manually labeling the class labels and bounding-boxes for large-scale datasets is a very expensive and tedious job, which is not cost-effective in practical applications. As a remedy, semi-supervised [38,48] and self-supervised [28] OD algorithms, which aim to employ the large amount of unlabeled data to improve the performance of OD, have been attracting much attention in recent years. In this paper, we focus on the semi-supervised objection detection (SSOD) methods.

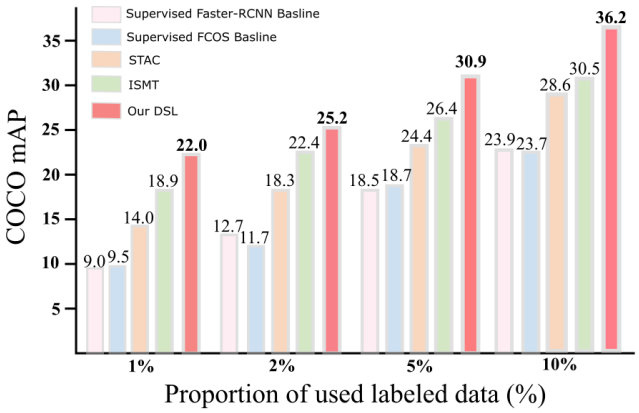

Figure 1. The SSOD performance comparisons between the proposed anchor-free based DSL and anchor-based methods STAC [38] and ISMT [48]. One can observe that anchor-based detector Faster-RCNN [36] and anchor-free based detector FCOS [44] have similar baseline performance under the supervised settings, while our proposed DSL achieves the state-of-the-art SSOD performance, outperforming the existing methods by a large margin.

The current state-of-the-art SSOD methods are pseudo-label based approaches [31, 38, 48, 51], while most of them are based on a two-stage anchor-based detector such as Faster-RCNN [36]. Specifically, they first use a teacher model to generate pseudo-labels for unlabeled images and then train a two-stage anchor-based detector with both labeled and unlabeled images. However, in real-world applications, the one-stage anchor-free based detectors (*e.g.*, FCOS [44]) are more attractive and practical since they are much easier and efficient to be deployed on resource limited devices without heavy pre/post-processing except NMS. Different from Faster-RCNN, the learning of FCOS is established on dense feature predictions; that is, each pixel is directly supervised by the corresponding label. Without the help of predefined anchors and multiple refinements of the predictions, the learning of anchor-free based detectors requires more careful guidance, especially under the SSOD settings. Unfortunately, few works on anchor-free SSOD

have been reported, and how to handle the dense pseudo-labels predicted by anchor-free detectors remains a challenging problem.

To address the above mentioned challenges, in this paper we propose a DenSe Learning (DSL) algorithm for anchor-free SSOD [1]. Specifically, to perform careful label guidance for dense learning, we first present an Adaptive Filtering (AF) strategy to partition pseudo-labels into three fine-grained parts, including background, foreground, and ignorable regions. Then we refine these pseudo-labels by using a MetaNet so as to remove the classification false-positives, which have higher prediction scores but are actually false predictions in category. Considering that the correctness of pseudo-labels determines the performance of SSOD models, we introduce an Aggregated Teacher (AT) to further enhance the stability and quality of the estimated pseudo-labels. Moreover, to improve the model generalization capability, we learn from shuffled image patches and regularize the uncertainty of dense feature maps to make them consistent among image scales. The main contributions of this paper are summarized as follows:

- A simple yet effective DenSe Learning (DSL) method is developed to improve the utilization of large-scale unlabelled data for SSOD. To our best knowledge, this is the first anchor-free method for SSOD.

- An Adaptive Filtering (AF) strategy is proposed to assign fine-grained pseudo-labels to each pixel; an Aggregated Teacher (AT) is introduced to enhance the stability and quality of estimated pseudo-labels; and learning from shuffled patches and uncertainty-consistency-regularization among scales are employed to improve the model generalization performance.

Extensive experiments conducted on MS-COCO [27] and PASCAL-VOC [8] demonstrate that the proposed DSL method achieves significant performance improvements over existing state-of-the-art SSOD methods.

2. Related Work

Semi-Supervised Learning for Image Classification. Recently, semi-supervised learning (SSL) has achieved significant progress in image classification with the rapid development of deep learning techniques. SSL aims to employ a large amount of unlabeled data to learn robust and discriminative classification boundaries. Specifically, self-ensembling is used in [19] to stabilize the learning targets for unlabeled data. A new measure of local smoothness of the conditional label distribution is proposed in [32] for improving the SSL learning performance. Mean teacher is employed in [42] to produce accurate labels instead of label ensembles. Generally speaking, the above consistency-based methods apply perturbations to the input image and then minimize the differences between their output predictions. These methods have proved to be effective at smoothing the feature manifold, and consequently improving the generalization performance of models. There are also some other techniques targeting at utilizing the unlabeled data to improve image classification, including self-training [6, 20, 23, 46], data augmentation [2, 37] and so on.

Though many SSL methods have been proposed for image classification, it is not a trivial work to transfer them to the task of object detection due to the complex architectural design and multi-task learning (classification and regression) nature of object detectors.

Object Detection is a fundamental task in computer vision. Current CNN-based object detectors can be categorized into anchor-based and anchor-free methods. Faster R-CNN [36] is a well-known and representative two-stage anchor-based detector. It consists of a region proposal network (RPN) and a region-wise prediction network (R-CNN) for detecting objects. Many works [1, 3, 4, 21, 24, 43] have been proposed to improve the performance of Faster RCNN. For anchor-free object detection, the state-of-the-art methods [13, 18, 30, 35, 44] mostly regard the center (*e.g.*, the center point or part) of an object as a foreground to define positives, and then predict the distances from positives to the four sides of the object bounding box (BBox). For example, FCOS [44] takes all the pixels inside the BBox as positives, and uses these four distances and a centerness score to detect objects. CSP [30] defines only the center point of the object box as positive to detect pedestrians with fixed aspect ratio. FoveaBox [18] regards pixels in the middle part of object as positives and learns four distances to perform detection. Without the need to set anchors, anchor-free detectors are much easier and more flexible to be deployed in real applications.

Semi-Supervised Object Detection (SSOD). SSOD aims to improve the performance of object detectors by using larger-scale unlabeled data. Since the manual annotation of object labels is very expensive, producing pseudo-labels for unlabeled data is very attractive. In [34, 39, 52], the pseudo-labels are produced by ensembling the predictions from different data augmentations. STAC [38] uses both weak and strong augmentations for model training, where strong augmentations are only applied to unlabeled data while weak augmentations are used to produce stable pseudo-labels. UBA [31] employs the EMA teacher [42] for producing more accurate pseudo-labels. ISMT [48] fuses the current pseudo-labels with history labels via NMS, and uses multiple detection heads to improve the accuracy of pseudo-labels. Instant-Teaching [51] combines more powerful augmentations like Mixup and Mosaic into the train-

[1] In this paper, we employ FCOS [44] as our baseline detector.

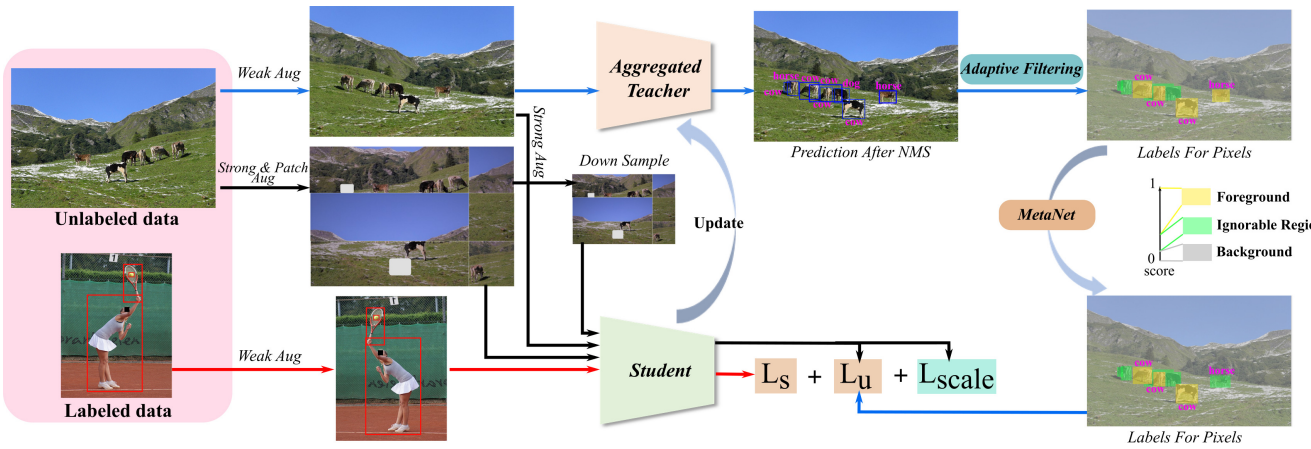

Figure 2. The pipeline of our proposed DenSe Learning (DSL) based SSOD method. The training data contain both labeled and unlabeled images. During each training iteration, a teacher model is employed to produce pseudo-labels for weakly augmented unlabeled images. In anchor-free based detectors like FCOS [44], each spatial location of the dense predictions will be assigned with one label, and the model performance is sensitive to noisy pseudo-labels. To alleviate this problem, an Adaptive Filtering strategy is proposed to split the pseudo-labels into three types, including background, foreground and ignorable regions. Moreover, there exist some false positive cases, which have higher scores but are obviously wrong predictions. Thus, a MetaNet is proposed to refine these cases. To improve the model generalization capability, unlabeled images are patch-shuffled and consistency regularizations are applied on these images with different scales. For improving the stability and quality of pseudo-labels, the teacher model is updated by the student models via aggregation, called Aggregated Teacher. After obtaining the fine-grained pixel-wise pseudo-labels, the detector can be optimized by the final loss, which is the sum of L_s, L_u and L_{scale}.

ing stage. Humble-Teacher [41] uses plenty of proposals and soft pseudo-labels for the unlabeled data. Certainty-aware pseudo-labels are tailored in [22] for object detection. E2E [47] uses a soft teacher mechanism for training with the unlabeled data. Almost all the above methods are built upon anchor-based detectors, *e.g.*, Faster RCNN, which are not convenient to deploy in real applications with limited resources. Therefore, in this work we develop, for the first time to our best knowledge, an anchor-free SSOD method.

3. Methods

3.1. Preliminary

For the convenience of expression, we first provide some notations for the SSOD task. Suppose that we have two sets of data, a labeled set $\mathcal{X} = \{X_i|_{i=1}^{N_l}\}$ and an unlabeled set $\mathcal{U} = \{U_i|_{i=1}^{N_u}\}$, where N_l and N_u are the number of labeled and unlabeled images, respectively, and $N_u \gg N_l$. Each labeled image has annotations of category $p^* \in [0, C-1]$ (C is the number of foreground classes) and annotations of bounding box (BBox) t^*. In an image, each region annotated by BBox and class label is called an instance. Without loss of generality, we take the anchor-free FCOS [44] detector as our baseline, which is composed of a ResNet50 [9] backbone, an FPN [26] neck and a dense head. To use both labeled and unlabeled data for training, the overall loss can be defined as follows:

$$L = L_s + \alpha L_u \tag{1}$$

where L_s and L_u denote supervised loss and unsupervised loss, respectively, and α is the hyper-parameter to control the contribution of unlabeled data.

Both of the supervised and unsupervised losses are normalized by the corresponding number of positive pixels in each mini-batch as follows:

$$L_s = \frac{1}{N_{pos}} \sum_i \sum_{h,w} (L_{cls}(X_{i,h,w}) + \mathbb{1}_{\{p^*_{h,w} \in [0,C-1]\}} L_{reg}(X_{i,h,w}) + \mathbb{1}_{\{p^*_{h,w} \in [0,C-1]\}} L_{center}(X_{i,h,w})) \tag{2}$$

$$L_u = \frac{1}{N_{pos}} \sum_i \sum_{h,w} (L_{cls}(U_{i,h,w}) + \mathbb{1}_{\{\bar{p}^*_{h,w} \in [0,C-1]\}} L_{reg}(U_{i,h,w}) + \mathbb{1}_{\{\bar{p}^*_{h,w} \in [0,C-1]\}} L_{center}(U_{i,h,w})) \tag{3}$$

where N_{pos} means the number of positive pixels in one mini-batch, $X_{i,h,w}$ means the predicted vector at spatial location (h, w) from the i^{th} image, $\bar{p}^*_{h,w}$ is the corresponding estimated pseudo-labels at location (h, w). L_{cls}, L_{reg} and L_{center} are the default losses used in FCOS [44]. $\mathbb{1}_{\{\cdot\}}$ is the indicator function, which outputs 1 if condition $\{\cdot\}$ is satisfied and 0 otherwise.

In this paper, we propose a ***DenSe Learning*** (DSL) algorithm for bridging the gap between SSOD and anchor-free detector. The pipeline of our DSL method is illustrated in Figure 2. It is mainly composed of an Adaptive Filtering (AF) strategy, a MetaNet, an Aggregated Teacher (AT) and an Uncertainty-Consistency regularization term, which are introduced in detail in the following sections.

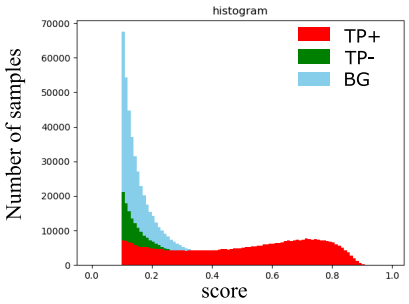

Figure 3. The distributions of TP+, TP- and BG when using 10% labeled data on COCO. 'TP+' means that the estimated instance has the same class ID as the ground-truth (GT) and the IOU of BBox is above 0.5. 'TP-' means that the estimated instance has the same class ID as GT but the IOU of BBox is below 0.5. 'BG' means that the estimated instance belongs to the background or has wrong class ID.

3.2. Adaptive Filtering Strategy

The FCOS [44] detector reduces the dependency on predefined anchors by introducing dense pixel-wise supervision. Though this is helpful for the easy deployment in actual applications, the performance of the model is sensitive to the quality of pixel-wise labels. Because the predicted pseudo-labels in SSOD will have noise no matter how powerful the detector is, the pixel-wise supervision for FCOS should be treated prudently. To this end, we propose an Adaptive Filtering (AF) strategy to elaborately handle the pseudo-labels for dense learning.

To exploit the unlabeled data, we need to assign a pseudo-label for each pixel in the output dense tensor. As shown in Figure 3, however, we can see that the TP+, TP- and BG instances coexist with each other, and their distributions are much more complex. If we simply use a single threshold to define foreground and background, many instances will be assigned with wrong labels, resulting in heavy noise and damaging the learning of an accurate detector. For example, if we set a relatively higher threshold 0.4 to define the positive instances, there will be many TP+ and TP- wrongly assigned to the background. Conversely, if we set a relatively lower threshold 0.1 to define the background instances, there will be many BG instances wrongly assigned to the foreground. Therefore, we propose to use multiple thresholds $\{\tau_1, \tau_2\}$ to partition the estimated instances into three parts: background, ignorable region and foreground:

$$\bar{p}^*_{h,w} = \begin{cases} Foreground : [0, \cdots, C-1] & p_{h,w} >= \tau_2, \\ Ignorable\ Region : [-1] & \tau_1 < p_{h,w} < \tau_2, \\ Background : [C] & p_{h,w} <= \tau_1. \end{cases} \tag{4}$$

where $p_{h,w}$ is the predicted score at location (h, w) (If not specified, it is the product of classification score and centerness score), and $\bar{p}^*_{h,w}$ is the corresponding pseudo-label.

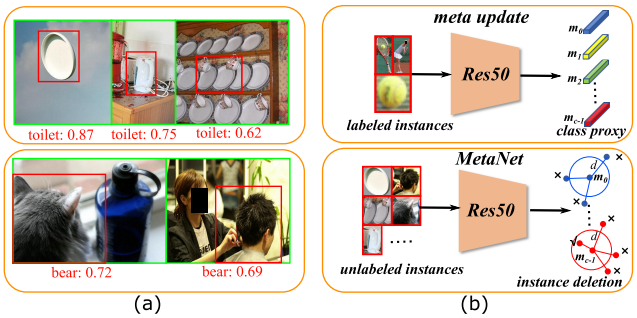

Figure 4. (a) The estimated classification-false-positive instances which have high scores but are obvious false predictions in category. (b) Our proposed MetaNet for refining the pseudo-labels of instances. '$\surd$' and '$\times$' mean reservation and deletion, *resp*.

Different from foreground and background regions, we ignore the gradients computation and propagation for ignorable regions as:

$$L_u = \frac{1}{N_{pos}} \sum_i \sum_{h,w} (\mathbb{1}_{\{\bar{p}^*_{h,w} \geq 0\}} L_{cls}(U_{i,h,w}) + \mathbb{1}_{\{\bar{p}^*_{h,w} \in [0, C-1]\}} L_{reg}(U_{i,h,w}) + \mathbb{1}_{\{\bar{p}^*_{h,w} \in [0, C-1]\}} L_{center}(U_{i,h,w})). \tag{5}$$

τ_1 in Eq. 4 is used to filter out the background and thus it is relatively easy to set. We set $\tau_1 = 0.1$ throughout our experiments. τ_2 is employed to filter out the foreground and it is harder to set for different classes. We propose to use a class-adaptive τ_2^k instead of a fixed τ_2:

$$\tau_2^k = (\frac{\sum_{h,w} \mathbb{1}_{\{\bar{p}^*_{h,w} == k\}} p_{h,w}}{N_{pos}})^{\beta} \tau, \tag{6}$$

where τ_2^k is the threshold for the k^{th} class, $\beta = 0.7$ is used to control the degree of focus on tail-classes, and $\tau = 0.35$ is used as a fixed reference threshold.

Remarks: Different from those anchor-based detectors, anchor-free detectors will predict each pixel as either background or foreground, and compute gradients for all of them. However, for unlabeled data, instances with scores within interval $[\tau_1, \tau_2^k]$ are noisy and confusing, and treating them as either foreground or background will degrade the detection performance. Therefore, in anchor-free SSOD we should explicitly set multiple fine-grained thresholds to identify not only the background and foreground but also the ignorable regions. The proposed AF strategy can well handle this problem and assign fine-grained and multi-level labels to the dense pixels, as illustrated in Figure. 2. We experimentally demonstrate that the AF strategy is very important for anchor-free SSOD.

3.3. MetaNet

Though AF has the ability to improve the quality of pseudo-labels for dense learning, there still exist some classification-false-positive instances, which have high

scores but are obvious false predictions, as shown in Figure 4(a). In order to handle these instances, we resort to using a MetaNet, as shown in Figure 4(b). We use a ResNet50 to implement the MetaNet. Before DSL training, we first pass all the labeled instances into the MetaNet and compute the following class-wise proxies m_k:

$$m_k = \frac{\sum_i f_{i,k}}{N_k}, \tag{7}$$

where $f_{i,k}$ is the 1-D feature vector of the i^{th} instance belonging to the k^{th} class, N_k is the number of instances of the k^{th} class. After obtaining the class-wise proxies, we refine the pseudo-labels by computing the cosine distance between the feature vector of the unlabeled instance and the corresponding class proxy vector. If the distance is smaller than a threshold $d = 0.6$, we will change the label 'Foreground' of this instance to the label 'Ignorable Region'.

Remarks: MetaNet is employed to rectify the predicted foreground class labels of those error-prone instances. It only performs the meta update step and thus can work in a plug-and-play manner. The computation of MetaNet only involves the class proxy update on the labeled instances without gradient back-propagation, and thus it is fast and the cost is negligible compared with the training of DSL. With the help of stable class proxies, we can successfully remove many classification-false-positive instances.

3.4. Aggregated Teacher

In pseudo-label based methods, the stability and quality of the predicted pseudo-labels are important to the final performance. Therefore, almost all the existing anchor-based methods [22, 31, 41, 47, 48] employ an EMA Teacher to improve the quality of pseudo-labels for the unlabeled data. As illustrated in Figure 5(a), EMA is usually performed in following manner:

$$\theta'^t = \epsilon\theta'^{t-1} + (1-\epsilon)\theta^t, \tag{8}$$

where ϵ is a smoothing hyperparameter, t means the iteration, θ and θ' are parameters of the student and teacher models, respectively.

EMA update aims to obtain a more stable and powerful teacher model via the ensemble of students. However, such an update in Eq. 8 might still be coarse and weak because it only aggregates parameters in the same layer at different iterations, without considering the correlation across layers. To further enhance the capability of teacher model, motivated by the dense aggregation mechanism [12, 49, 50], we introduce an Aggregated Teacher (AT), which performs not only parameter aggregation across time but also recurrent layer aggregation across layers, as illustrated in Figure 5(b). Specifically, for parameter aggregation, we still adopt the existing EMA update as in Eq. 8. While for layer aggregation, to avoid the problem of heavy parameter, we follow

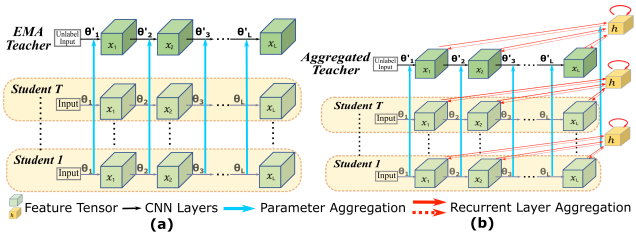

Figure 5. The illustration of (a) EMA Teacher and (b) our Aggregated Teacher. EMA teacher performs aggregation only over parameters, while our Aggregated teacher performs aggregation over both parameters and layers.

the recurrent learning [11, 25, 50] and use a recurrent layer aggregation mechanism as bellow:

$$x_{l+1} = \theta_{l+1}[x_l + h_l] + x_l, \tag{9}$$

$$h_{l+1} = g_2[g_1[\theta_{l+1}[x_l + h_l]] + h_l], \tag{10}$$

where x_l is the l^{th} layer's tensor in CNN and θ_l denotes the corresponding convolution parameters. h_l is the hidden state tensor for the l^{th} layer, and h_1 is initialized with zero. g_1 and g_2 are the corresponding 1×1 and 3×3 Conv layers used for recurrent computing, which are parameter-shared across the adjacent layers within the same stage. $*[\cdot]$ indicates the convolution operation between input tensor '$\cdot$' and parameter '$*$'. By using the recurrent mechanism, the number of introduced parameters is negligible. One can see from Eq. 9 that it will degrade to the default residual unit of ResNet when the hidden state h_{l-1} is removed. In other words, the recurrent layer aggregation can be easily applied to the current residual CNN models. Moreover, since neck and heads in the detector are very shallow, we only perform layer aggregation over the backbone.

Remarks: Since the parameter aggregation in EMA Teacher treats each layer independently, the relationship between layers might be destroyed during aggregation, and thus one aggregated layer may not work well with the adjacent ones. Therefore, layer aggregation is considered in our model. By explicitly using the hidden state to connect the current layer with the previous layers, the knowledge propagation will be more stable and accurate. Moreover, the shared recurrent layers impose regularization over the propagated information. Compared with EMA Teacher, the Aggregated Teacher is able to produce more stable and accurate pseudo-labels for dense learning.

3.5. Uncertainty Consistency

By using the proposed AF, MetaNet and AT, the dense pixel-wise pseudo-labels can be obtained to supervise the learning of SSOD models by optimizing the loss L_u. In order to further improve the generalization capability of the SSOD model, we propose to regularize the uncertainty consistency over the unlabeled images. From Figure 6, one

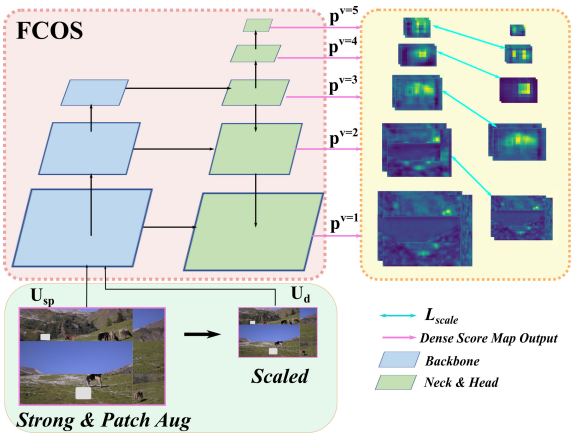

Figure 6. Illustration of the uncertainty consistency regularization among scales. The input images come from the same unlabeled image U_i.

can see that the input consists of a pair of images: Strong & Patch Augmented image (U_{sp}) and the corresponding Down-sampled image (U_d). The downsampling ratio is set to $r = 2$ in producing U_d. By patch shuffle augmentation, we randomly crop an image into several parts along the horizontal or vertical directions and then shuffle these parts (detailed algorithm can be found in Algorithm 1). Both the two images will be fed into our detector, producing dense score maps at different scale levels. (In FCOS, there are 5 levels, *i.e.*, $v \in [1, \cdots, 5]$.)

To improve the generalization performance of SSOD, we adopt the following regularization loss:

$$L_{scale} = \sum_{v=1}^{4} \|p^v[U_d] - p^{v+1}[U_{sp}]\|_2^2, \qquad (11)$$

where $p^v[U_*]$ indicates the score map p^v derived from image U_*. Since the downsampling ratio $r = 2$, $p^v[U_d]$ has the same resolution as $p^{v+1}[U_{sp}]$, and they are constrained to be consistent.

Remarks: The output dense score maps reveal the uncertainty or the reliability of the predicted label for each pixel. The lower the score is, the higher the uncertainty that the pixel belongs to a foreground object. Data uncertainty has been widely used to indicate the data importance in previous works [6, 10, 15, 16, 45]. In this paper, we regularize the uncertainty consistency. Patch shuffle is used to reduce the dependency of foreground objects on their surrounding contexts, improving the model robustness to context variations. In addition, to ensure consistent outputs among scales, L_{scale} is then defined to improve the model robustness to object scaling variations.

By far, all the components of our DSL have been described, and the overall pipeline is shown in Figure 2.

Algorithm 1: Patch Shuffle

Input: Unlabeled image U;
Output: Patch shuffled image U_p;
Initialization: $U^0 = U$, total iteration number J;
for $j = 0, \cdots, J-1$ **do**
  (1) Mode m: randomly select a mode from ['horizontal','vertical'];
  (2) Normalized size s: randomly generate s from interval $[0, 1]$;
  (3) Crop U^j into two parts based on mode m and normalized size s;
  (4) Shuffle the order of the two parts, and concatenate them into a new image $\hat{U}^j$;
  (5) $U^{j+1} = \hat{U}^j$;
end

4. Experiments

Datasets & Evaluation Metrics: We conduct experiments on the popular object detection benchmarks, including MS-COCO [27] and PASCAL-VOC [8]. MS-COCO contains more than $118k$ labeled images, and there are about $850k$ instances from 80 classes. In addition, there are $123k$ unlabeled images provided for semi-supervised learning. VOC07 contains 5,011 training images from 20 classes, while VOC12 has 11,540 training images.

On MS-COCO, we follow the settings in STAC [38] and evaluate with both the protocols of Partially Labeled Data and Fully Labeled Data. The former randomly samples 1%, 2%, 5% and 10% of the training data as labeled data, and treats the remainder as unlabeled data. (For this protocol, we create 3 data folds and report the mean results over them.) The latter uses all the training data as labeled data and the additional unlabeled data as unlabeled samples. We adopt the mean average precision $AP_{50:90}$ (denoted by mAP) as the evaluation metric.

For experiments on PASCAL-VOC07, following STAC [38], we use the VOC07 training set as the labeled data, and the VOC12 training set or together with the images from the same 20 classes in MS-COCO (denoted by COCO20) as the unlabeled data. We adopt VOC default AP_{50} metric and COCO default mAP metric as the evaluation metrics.

Implementation Details: We adopt the popular anchor-free detector FCOS [44] with ResNet50 [9] as backbone, and FPN [27] as neck and dense heads. Images in MS-COCO are resized to have shorter edge 800, or 640 if the longer edge is less than 1,333. Images in PASCAL-VOC are resized to have shorter edge 600, or 480 if the longer edge is less than 1,000. For fair comparison, following [31, 38], in all experiments, random flip is used as weak augmentation, while strong augmentation includes random flip, color jittering and cutout. The iteration J is set to 2 in Patch Shuffle. For training configurations, learning rate starts from

Table 1. The mAP performance (%) of competing methods on the MS-COCO [27] dataset. The used protocol is Partially Labeled Data. † means that the method uses a larger batch size 32 or 40, and ‡ indicates that strong augmentation is applied on the labeled data. Note that †, ‡ are not the default settings in STAC [38] but they will improve the performance of both supervised baseline and SSOD. 'Supervised' means that only the corresponding labeled data are used for training, and this is set as the baseline for SSOD.

Methods		Deployment	1%	2%	5%	10%
Anchor-based	Supervised [38]	Hard	9.05 ± 0.16	12.70 ± 0.15	18.47 ± 0.22	23.86 ± 0.81
	CSD [14]	Hard	11.12 ± 0.15	14.15 ± 0.13	18.79 ± 0.13	24.50 ± 0.15
	STAC [38]	Hard	13.97 ± 0.35	18.25 ± 0.25	24.38 ± 0.12	28.64 ± 0.21
	IT [51]	Hard	16.00 ± 0.20	20.70 ± 0.30	25.50 ± 0.05	29.45 ± 0.15
	ISMT [48]	Hard	18.88 ± 0.74	22.43 ± 0.56	26.37 ± 0.24	30.53 ± 0.52
	Humble [41]	Hard	16.96 ± 0.38	21.72 ± 0.24	27.70 ± 0.15	31.60 ± 0.28
	UB† [31]	Hard	20.75 ± 0.12	24.30 ± 0.97	28.27 ± 0.11	31.50 ± 0.10
	E2E†‡ [47]	Hard	20.46 ± 0.39	-	30.74 ± 0.08	34.04 ± 0.14
Anchor-free	Supervised(Ours)	Easy	9.53 ± 0.23	11.71 ± 0.26	18.74 ± 0.18	23.70 ± 0.22
	DSL(Ours)	Easy	**22.03** ± 0.28	**25.19** ± 0.37	**30.87** ± 0.24	**36.22** ± 0.18

0.01 and is divided by 10 at 16 and 22 epochs. The max epoch is 24. α is set to 3 and 1 for the partially and fully labeled protocols, *resp*, and 2.5 for VOC. ϵ is set to 0.99. For parameter τ_2^k, we set it within the range $[0.25, 0.35]$. All of our experiments are based on Pytorch [33] and MMDetection [7]. We use 8 NVIDIA-V100 GPUs with 32G memory per GPU. For each GPU, we randomly sample 2 images from labeled set and unlabeled set with ratio 1:1.

4.1. Comparison with State-of-the-Arts

We compare the proposed DSL with existing SOTA methods that are based on anchor-based detectors such as Faster-RCNN [36] and SSD [29]. The results are shown in Tables 1, 2 and 3.

From Table 1, one can see that under the supervised setting of the Partially Labeled Data protocol in COCO, our anchor-free detector achieves similar baseline performance to those anchor-based detectors, *i.e.*, 9.53 vs. 9.05, 11.71 vs. 12.70, 18.74 vs. 18.47 and 23.7 vs. 23.86 with 1%, 2%, 5% and 10% labeled data, respectively. This means that anchor-free and anchor-based SSOD models are comparable when partially labeled data are used. After applying the proposed DSL algorithm, the SSOD performance can be significantly and consistently improved over the baselines under all protocols. DSL outperforms all the competing methods by a large margin, demonstrating the effectiveness and superiority of our method.

We also conduct experiments following the Fully Labeled Data protocol of COCO. The results are shown in Table 2. Since the reported performance of those supervised methods varies a lot in the original works, we report their results together with their baselines, and compare their relative performance improvements. From Table 2, one can see that our DSL achieves the largest performance improvement, *i.e.*, 3.6 mAP gain. The results on PASCAL-VOC are listed in Table 3. We can see that the proposed DSL also achieves significant performance improvements over the supervised baselines as well as all the compared methods.

Table 2. The mAP performance (%) of competing methods on the MS-COCO [27] dataset. The used protocol is Fully Labeled Data.

Methods		Deployment	100%
Anchor-based	STAC [38]	Hard	37.6 $\xrightarrow{1.6}$ 39.2
	ISMT [48]	Hard	37.8 $\xrightarrow{1.8}$ 39.6
	UB† [31]	Hard	40.2 $\xrightarrow{1.1}$ 41.3
	E2E†‡ [47]	Hard	40.9 $\xrightarrow{3.6}$ 44.5
Anchor-free	DSL(Ours)	Easy	**40.2** $\xrightarrow{\mathbf{3.6}}$ **43.8**

In summary, the results in Tables 1, 2 and 3 all demonstrate the effectiveness of our DSL method. It is worth mentioning that the proposed DSL is much easier to be deployed in real applications due to its negligible pre/post-processing costs compared to anchor-based methods, showing the great potential values of the anchor-free SSOD algorithm.

4.2. Ablation Studies

To better understand how the proposed DSL works, we conduct a series of ablation studies under the MS-COCO 10% labeled data protocol.

Effectiveness of each component. The contributions of different components of DSL are listed in Table 4. From this table, one can see that by using AF, the performance can be significantly improved from 23.7 to 32.2 mAP, which has already surpassed most SOTA methods shown in Table 1. By adopting the MetaNet to refine the foreground pseudo-labels, the performance can be further improved to 32.5. By applying AT to encourage the stability and quality of the pseudo-labels, the performance is further improved to 34.5 mAP. Finally, by learning from shuffled patches and constraining the consistency among image scales, the overall model becomes more robust and exhibits higher accuracy, *i.e.*, 36.2 mAP. The ablation studies in Table 4 verify the effectiveness of each module in DSL.

Ablation studies on AF. Table 5 shows the ablation studies on our AF strategy. In order to demonstrate the importance of multiple thresholds, we experiment with a

Table 3. The results (%) of competing methods on the PASCAL-VOC [8] dataset. The performances are evaluated on the VOC07 test set.

Methods		Deployment	Unlabeled: VOC12		Unlabeled: VOC12 + COCO20	
			AP_{50}	$AP_{50:90}$	AP_{50}	$AP_{50:90}$
Anchor-based	Supervised [38]	Hard	72.75	42.04	72.75	42.04
	CSD [14]	Hard	74.7	-	75.1	-
	STAC [38]	Hard	77.45	44.64	79.08	46.01
	IT [51]	Hard	78.3	48.7	79	49.7
	ISMT [48]	Hard	77.23	46.23	77.75	49.59
	UB† [31]	Hard	77.37	48.69	78.82	50.34
Anchor-free	Supervised(Ours)	Easy	69.6	45.9	69.6	45.9
	DSL (Ours)	Easy	**80.7**	**56.8**	**82.1**	**59.8**

Table 4. Effectiveness of each component of the proposed DSL method. '+' means training by the proposed method.

Methods	mAP
Supervised	23.7
+ AF	32.2
+ MetaNet	32.5
+ AT	34.5
+ Patch-Shuffle	34.9
+ L_{scale}	36.2

Table 5. Ablation studies on Adaptive Filtering.

Methods	Single threshold				AF(fixed τ_2^k)			AF
	0.05	0.1	0.2	0.3	0.2	0.3	0.4	
mAP	27.1	28.8	30.7	27.5	34.3	36.0	35.6	**36.2**

Table 6. Ablation studies on Aggregated Teacher. 'LA' means layer aggregation.

Methods	No teacher	+ EMA	+ LA	AT
mAP	33.0	34.1	35.0	**36.2**

Table 7. Ablation studies on loss weight α for unlabeled data. 'fail' means that the training loss will easily get to 'nan'.

α	1	2	3	4
mAP	33.9	35.4	**36.2**	fail

single threshold strategy as reference, where instances are regarded as foreground if their scores are above the threshold and background otherwise. One can see that the single threshold strategy cannot achieve satisfactory performance. The best result is only 30.7 mAP when the threshold is set to 0.2, indicating that there are many instances being wrongly defined by a single threshold. In contrast, by using our multi-level thresholds strategy, *i.e.*, AF, the performance can be significantly improved: even by using a fixed τ_2^k=0.3, the result can be improved to 36.0 mAP; and when the adaptive τ_2^k is used for each class, it can be further improved to 36.2 mAP, showing the effectiveness and importance of our AF strategy.

Ablation studies on AT. From Table 6, one can see that layer aggregation (LA) achieves higher performance gain than EMA because it considers the fine-grained relationships across layers, while EMA just simply aggregates layer-wise parameters independently so that the relationships between layers can be harmed. In addition, by employing both EMA and LA, our AT can further improve the performance to 36.2 mAP. This implies that aggregations over parameters and layers are actually complementary.

Ablation studies on loss weight α. From Table 7, one can see that the performance peaks around $\alpha = 3$. A too large weight such as $\alpha = 4$ will give the model too many chances to employ the unlabeled images in training, and hence reduce the stability of the model.

Discussions. In anchor-based SSOD, the negative/ignorable instances have been implicitly handled by label assigner and sampler, and we only need to consider how to recall the foreground instances via a threshold. In contrast, in anchor-free SSOD the multi-level pseudo-labels should be explicitly considered due to the pixel-wise gradient propagation. This can be demonstrated by our AF strategy as in Table 5. Moreover, without the help of predefined anchors for scale variances, FPN [27] with a dense head has been widely used in anchor-free detectors to address the scaling issue. Thus L_{scale} can be generally adopted and regarded as a default trick in anchor-free SSOD, and this is verified to be effective in Table 4. In summary, most of our techniques are proposed by considering the special characteristics of anchor-free detectors, and our work in this paper makes the first step towards anchor-free SSOD.

5. Conclusion

In this paper, we made the first attempt, to the best of our knowledge, to bridge the gap between SSOD and anchor-free detector, and developed a DSL based SSOD method. The DSL was built upon several novel techniques, such as Adaptive Filtering, Aggregated Teacher and uncertainty regularization. Our experiments showed that the proposed DSL outperformed the state-of-the-art SSOD methods by a large margin on both COCO and VOC datasets. It is expected our work can inspire more and in-depth explorations on anchor-free SSOD methods.

References

[1] Sean Bell, C Lawrence Zitnick, Kavita Bala, and Ross Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2874–2883, 2016. 2

[2] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*, 2019. 2

[3] Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *European conference on computer vision*, pages 354–370. Springer, 2016. 2

[4] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 2

[5] 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. 1

[6] Binghui Chen and Weihong Deng. Weakly-supervised deep self-learning for face recognition. In *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2016. 2, 6

[7] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 7

[8] 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, 6, 8

[9] 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*, pages 770–778, 2016. 3, 6

[10] Jay Heo, Hae Beom Lee, Saehoon Kim, Juho Lee, Kwang Joon Kim, Eunho Yang, and Sung Ju Hwang. Uncertainty-aware attention for reliable interpretation and prediction. *arXiv preprint arXiv:1805.09653*, 2018. 6

[11] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 5

[12] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 5

[13] Lichao Huang, Yi Yang, Yafeng Deng, and Yinan Yu. Densebox: Unifying landmark localization with end to end object detection. *arXiv preprint arXiv:1509.04874*, 2015. 2

[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. 7, 8

[15] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *arXiv preprint arXiv:1703.04977*, 2017. 6

[16] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018. 6

[17] Kang Kim and Hee Seok Lee. Probabilistic anchor assignment with iou prediction for object detection. In *ECCV*, 2020. 1

[18] Tao Kong, Fuchun Sun, Huaping Liu, Yuning Jiang, Lei Li, and Jianbo Shi. Foveabox: Beyound anchor-based object detection. *IEEE Transactions on Image Processing*, 29:7389–7398, 2020. 2

[19] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 2

[20] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013. 2

[21] Hyungtae Lee, Sungmin Eum, and Heesung Kwon. Me r-cnn: Multi-expert r-cnn for object detection. *IEEE Transactions on Image Processing*, 29:1030–1044, 2019. 2

[22] Hengduo Li, Zuxuan Wu, Abhinav Shrivastava, and Larry S Davis. Rethinking pseudo labels for semi-supervised object detection. *arXiv preprint arXiv:2106.00168*, 2021. 3, 5

[23] Xinzhe Li, Qianru Sun, Yaoyao Liu, Qin Zhou, Shibao Zheng, Tat-Seng Chua, and Bernt Schiele. Learning to self-train for semi-supervised few-shot classification. *Advances in Neural Information Processing Systems*, 32:10276–10286, 2019. 2

[24] Yanghao Li, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Scale-aware trident networks for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6054–6063, 2019. 2

[25] Tsungnan Lin, Bill G Horne, Peter Tino, and C Lee Giles. Learning long-term dependencies in narx recurrent neural networks. *IEEE Transactions on Neural Networks*, 7(6):1329–1338, 1996. 5

[26] 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. 3

[27] 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, 6, 7, 8

[28] Songtao Liu, Zeming Li, and Jian Sun. Self-emd: Self-supervised object detection without imagenet. *arXiv preprint arXiv:2011.13677*, 2020. 1

[29] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 7

[30] Wei Liu, Shengcai Liao, Weiqiang Ren, Weidong Hu, and Yinan Yu. High-level semantic feature detection: A new perspective for pedestrian detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5187–5196, 2019. 2

[31] 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. 1, 2, 5, 6, 7, 8

[32] 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

[33] Pytorch. https://pytorch.org/. 7

[34] 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. 2

[35] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 2

[36] 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. 1, 2, 7

[37] 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. 2

[38] 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. 1, 2, 6, 7, 8

[39] Xiaolin Song, Binghui Chen, Pengyu Li, Biao Wang, and Honggang Zhang. Prnet++: Learning towards generalized occluded pedestrian detection via progressive refinement network. *Neurocomputing*, 2022. 2

[40] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14454–14463, 2021. 1

[41] 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. 3, 5, 7

[42] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems*, 30, 2017. 2

[43] Wanxin Tian, Zixuan Wang, Haifeng Shen, Weihong Deng, Yiping Meng, Binghui Chen, Xiubao Zhang, Yuan Zhao, and Xiehe Huang. Learning better features for face detection with feature fusion and segmentation supervision. *arXiv preprint arXiv:1811.08557*, 2018. 2

[44] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 1, 2, 3, 4, 6

[45] 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. 6

[46] 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. 2

[47] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. *arXiv preprint arXiv:2106.09018*, 2021. 3, 5, 7

[48] 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. 1, 2, 5, 7, 8

[49] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2403–2412, 2018. 5

[50] Jingyu Zhao, Yanwen Fang, and Guodong Li. Recurrence along depth: Deep convolutional neural networks with recurrent layer aggregation. *Advances in Neural Information Processing Systems*, 34, 2021. 5

[51] 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. 1, 2, 7, 8

[52] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. Rethinking pre-training and self-training. *Advances in Neural Information Processing Systems*, 33, 2020. 2