Object Detection in Densely Packed Scenes via Semi-Supervised Learning with Dual Consistency

Chao Ye[1*] , **Huaidong Zhang**[2*] , **Xuemiao Xu**[1,3,4,5†] , **Weiwei Cai**[1] , **Jing Qin**[2] and **Kup-Sze Choi**[2]
[1]South China University of Technology
[2]Centre for Smart Health, The Hong Kong Polytechnic University
[3]State Key Laboratory of Subtropical Building Science
[4]Ministry of Education Key Laboratory of Big Data and Intelligent Robot
[5]Guangdong Provincial Key Lab of Computational Intelligence and Cyberspace Information
alaylmyc@gmail.com, hd-wilson.zhang@polyu.edu.hk, xuemx@scut.edu.cn, cweiwei349@gmail.com,
{harry.qin, thomasks.choi}@polyu.edu.hk

Abstract

Deep neural networks have been shown to be very powerful tools for object detection in various scenes. Their remarkable performance, however, heavily depends on the availability of a large number of high quality labeled data, which are time-consuming and costly to acquire for scenes with densely packed objects. We present a novel semi-supervised approach to addressing this problem, which is designed based on a common teacher-student model, integrated with a novel intersection-over-union (IoU) aware consistency loss and a new proposal consistency loss. The IoU-aware consistency loss evaluates the IoU over the prediction pairs of the teacher model and the student model, which enforces the prediction of the student model to approach closely to that of the teacher model. The IoU-aware consistency loss also reweights the importance of different prediction pairs to suppress the low-confident pairs. The proposal consistency loss ensures proposal consistency between the two models, making it possible to involve the region proposal network in the training process with unlabeled data. We also construct a new dataset, namely RebarDSC, containing 2,125 rebar images annotated with 350,348 bounding boxes in total (164.9 annotations per image average), to evaluate the proposed method. Extensive experiments are conducted over both the RebarDSC dataset and the famous large public dataset SKU-110K. Experimental results corroborate that the proposed method is able to improve the object detection performance in densely packed scenes, consistently outperforming state-of-the-art approaches. Dataset is available in https://github.com/Armin1337/RebarDSC.

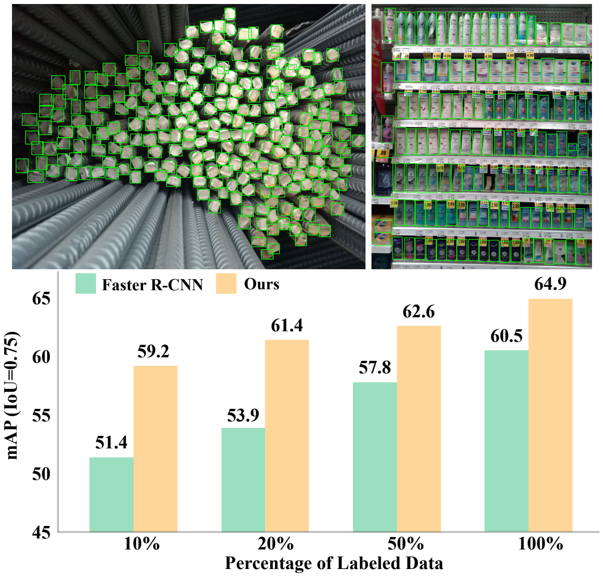

Figure 1: Top: labeled objects in densely packed scenes. Top left: the RebarDSC dataset used in this paper; Top right: the SKU-110K dataset. Bottom: qualitative comparison of Faster R-CNN and our semi-supervised learning method.

1 Introduction

Object detection has been widely explored under rapid development of deep neural networks for real-world scenes [Ren *et al.*, 2016; Lin *et al.*, 2017b]. Despite their remarkable performance, these deep detectors usually need enormous labeled data to learn from each specific scene. This requirement becomes an issue when heavy annotations are concerned, for example, in the images capturing densely packed objects. In fact, densely packed scenes generally contain hundreds or even more objects within one image, and it is time-consuming and costly for human annotators to label the data manually; see Figure 1 for example.

A solution to preclude the reliance on labeled data is to enable the network to learn from unlabeled data. This can be achieved by leveraging semi-supervised learning (SSL),

[*] Both authors contributed equally.
[†] Corresponding author: Xuemiao Xu.

which deals with situations where the training set consists of limited labeled data and enormous unlabeled data. Recent semi-supervised studies for object detection mainly harness consistency regularization [Jeong *et al.*, 2019; Zhao *et al.*, 2020; Hong *et al.*, 2020], which transforms the image data with augmentation to obtain augmented images, and then feeds the pair of images into the detection model, and finally optimizes the model to produce the same predictions. Consistency regularization has shown remarkable performance on many objection detection tasks [Gao *et al.*, 2019; Cai *et al.*, 2019; Zhao *et al.*, 2020], and we adopt it as the baseline in our study.

However, there are two significant challenges in consistent learning when it is applied in densely packed object detection. First, in most existing settings, the prediction consistency is evaluated over multiple pairs of object prediction, but, as the objects in densely packed scenes are very close to each other, it is difficult, or not impossible, to precisely construct pairs to align predictions of different objects. Second, the design of the consistency loss is crucial to the detection performance. Unfortunately, with many works focusing on sparse object detection, there is no suitable consistency loss for handling densely packed scenes.

In this paper, we propose a novel semi-supervised learning approach with a new consistency loss to tackle the challenging problem of accurately detecting objects from densely packed scenes. The proposed consistency loss is composed of two separate, yet complementary, components: an IoU-aware consistency loss and a proposal consistency loss. Specifically, our approach has two detectors, i.e. the teacher model and the student model. During training, we feed unlabeled images into the teacher network and the augmented images into the student network to obtain the predictions. We then align the student prediction and the teacher prediction with the IoU metric, so that each prediction box of the teacher model has an aligned box from the prediction of the student model. With the alignment results, we calculate the IoU-aware consistency loss between each pair of predictions. The IoU-aware consistency loss not only maximizes the IoU of paired predictions, but also reweights the other consistency losses with IoU as the weights. Furthermore, since the region proposal network (RPN), an extra network branch widely used in detectors, can not learn from the unlabeled data based on the prediction consistency loss due to undifferentiable gradients, we design a proposal consistency loss, which bridges the proposal results from the teacher model to the student model, enforcing the proposal results to be consistent. With the combination of the proposed losses, the detection network can learn from both the labeled data and unlabeled data simultaneously, further alleviating the reliance on the labeled data. To comprehensively evaluate the proposed approach, also considering the lack of datasets on densely packed scenes, we further construct a new dataset, namely RebarDSC, containing 2,125 rebar images annotated with 350,348 bounding boxes in total (164.9 annotations per image average). Our contributions can be summarized as follows:

- We present a novel semi-supervised learning approach with dual consistency losses for object detection in densely packed scenes.
- We employ two separate, yet complementary, consistency losses to drive the teacher-student model to learn both labeled and unlabeled data, one aiming at aligning IoU metric while the other attempting to align proposals.
- We build a new real-world dataset (RebarDSC) that contains more than 2K high-resolution rebar images with full annotations to enrich the dataset for densely packed scenes.

2 Related Work

2.1 Objection Detection in Densely Packed Scenes

Recently, densely packed object detection has drawn much attention. The existing research mainly adopts detection network [Ren *et al.*, 2016; Lin *et al.*, 2017b] as basic architecture, and tackles the unique challenges from densely packed scenes. For example, Hsieh *et al.* [2017] proposed a spatially regularized loss to learn neighbor cues within densely packed objects to improve the quality of proposals. Goldman *et al.* [2019] presented the large dataset SKU-110K collected in densely packed scenes, and proposed an EM merging unit to reduce the merging errors in non-maximum suppression (NMS). Similarly, Wang *et al.* [2020c] improved NMS performance and designed the HNMS algorithm that was much faster than NMS. Cai *et al.* [2020] developed a guided attention network to reserve the resolution of feature maps and learn integrated high-level features with supervised attention, so that small objects in dense scenes can also be detected. Pan *et al.* [Pan *et al.*, 2020] tackled the challenge of oriented objects by proposing a dynamic refinement network that could dynamically adjust the receptive fields. Chen *et al.* [2020] also handled the issue of oriented objects by proposing PIoU loss to evaluate the angle and IoU between ground truths and predictions. The above works focus on improving the detector performance based on fully supervised settings. On the contrary, we present here a semi-supervised method to reduce the need of labeled data in densely packed scenes.

2.2 Semi-supervised Learning

Recent semi-supervised learning methods mainly exploit consistency regularization to train with unlabeled data, for example Temporal Ensembling [Laine and Aila, 2016], Mixmatch [Berthelot *et al.*, 2019], Fixmatch [Sohn *et al.*, 2020a]. Here, we focus our review on the Mean Teacher [Tarvainen and Valpola, 2017], which we used in the study. Mean Teacher follows the pioneering work in [Laine and Aila, 2016], proposed to learn the prediction consistency between two models saved in different epochs. In [Tarvainen and Valpola, 2017], two models, the student model and the teacher model, were saved under the Mean Teacher framework. In their implementation, they considered the newly-updated model as the student, and the average of consecutive student models as the teacher. In this way, the student model could learn from unlabeled data based on the consistency losses, and the teacher model could be updated through the exponential moving average (EMA) after each iteration. In this paper, we construct the IoU-aware consistency loss and

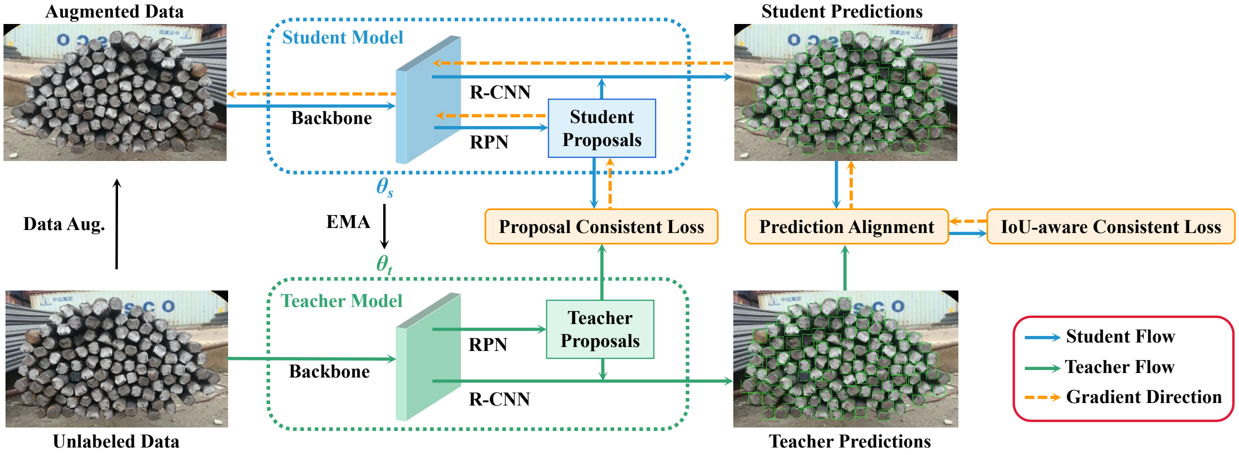

Figure 2: Overview of our semi-supervised learning framework for densely packed object detection with unlabeled data.

the proposal consistency loss based on consistency regularization, to establish the SSL framework for object detection in densely packed scenes.

Researches have been conducted recently to explore SSL for object detection [Jeong *et al.*, 2019; Sohn *et al.*, 2020b; Zhao *et al.*, 2020; Wang *et al.*, 2020a], which demonstrate good performance when compared with the methods using labeled data only. However, these methods are not designed for SSL in densely packed scenes. In this paper, experiments are conducted to compare our method with these methods.

3 Method

3.1 Overview

Problem Formulation. Given an image captured from densely packed scenes, our objective is to localize all the objects with bounding boxes. We train the detector under the semi-supervised setting, so that the network can learn from the unlabeled data. Formally, we have N_s training samples with dense object annotation $I_s = \{x_i, b_i\}_{i=1}^{N_s}$, and N_t training samples with image only $I_t = \{x_i\}_{i=1}^{N_t}$. Here x denotes the image data and b denotes the annotation set of each image sample.

Network Architecture. Figure 2 illustrates the architecture of our learning framework. The architecture is based on the design of Mean Teacher [Tarvainen and Valpola, 2017]. We have two detection models, the student model and the teacher model. They share the same network architecture but with different parameters. Here we use ResNet [He *et al.*, 2016] integrated with FPN [Lin *et al.*, 2017a] as the feature extractor, and Faster R-CNN [Ren *et al.*, 2016] as the detector head. We also test our method with different detector architectures to show that our method can perform well with diverse architectures. Refer to the experiments for details.

Learning Strategy. For the training with unlabeled data, the student model is learned from the consistency losses, which we have described the designs in Sec. 3.2 and Sec. 3.3. The teacher model does not directly learn from any loss, but updates its parameters with the EMA of the student model. Formally, we define θ_t as the teacher parameters, θ_s as the student parameters, and then update the teacher parameters after each iteration with:

$$\theta_t = (1 - \lambda_{ema})\theta_s + \lambda_{ema}\theta_{t-1}, \qquad (1)$$

where λ_{ema} denotes the decay factor.

Data Augmentation. Following the work in [Tarvainen and Valpola, 2017], augmentation should be performed on the input data of the student model and teacher model to avoid overfitting. Specifically, we fed the input images before augmentation to the teacher model, and the images with augmentation to the student model. The augmentation includes horizontal flips which are performed over every augmented image, brightness transformation which adds the brightness with probability 50%, and contrast transformation which multiplies the contrast with the value between $[0.9, 1.1]$ and probability 50%.

3.2 IoU-aware Consistency Loss

Prediction Alignment. Different from the task of image recognition of a single prediction per image, the prediction results of object detection here consist of multiple bounding boxes. Therefore, before calculating the prediction consistency loss, we first need to align the predictions of the teacher model and the student model into pairs. This step is more challenging under the densely packed scenes, since the objects are close to each other and have a similar appearance. Inspired by [Chen *et al.*, 2020; Pan *et al.*, 2020] that Intersection-over-Union (IoU) is a good metric to evaluate the bounding box accuracy, we propose to align the predictions with IoU. Formally, we denote the set of box predictions from the student model and the teacher model as $B_s = \{b_s\}$, and $B_t = \{b_t\}$ respectively. For each b_t^i from the teacher, we find the b_a^i with largest IoU from the set B_s to construct a pair, that is:

$$b_a^i = \underset{b_s}{\text{argmax}} \left(\text{IoU}(\, b_s, b_t^i \,) \right) \; . \qquad (2)$$

After alignment, we obtain the teacher predictions $B_t = \{b_t\}$ and its alignment predictions from the student $B_a = \{b_a\}$, where $|B_t| = |B_a|$. Then we calculate the IoU-aware consistency loss over all the pairs.

Loss Design

The recent works on generic object detection [Jiang *et al.*, 2018] and densely packed object detection [Chen *et al.*, 2020] have shown that maximizing the IoU between prediction and ground-truth can improve the detector under supervised learning. Inspired by these works, we introduce the IoU loss to semi-supervised learning. That is, we learn the unlabeled data by maximizing the IoU between prediction pairs from the two models. To this end, we present the IoU-aware consistent loss as:

$$\mathcal{L}_{iou} \;=\; \lambda_{iou} \frac{\sum_i^{|B_t|} \left\| 1 - \mathrm{IoU}(b_t^i, b_a^i) \right\|_2}{|B_t|}, \qquad (3)$$

where λ_{iou} denotes the weight of IoU-aware consistency loss. Unlike the recent SSL work [Wang *et al.*, 2020b] estimates the IoU with extra network layers, we estimate the IoU with the pair of regressed boxes after NMS directly.

To further stabilize the consistency regularization, we follow the recent work in [Zhao *et al.*, 2020] to calculate the center-aware consistency loss, the probability-aware consistency loss and the size-aware consistency. However, we find that many false alignments occur for densely packed scenes. If we directly apply these losses, the false pairs will ruin the model training. Hence, we further re-weight the gradient from each prediction pairs by their IoU, so that the pairs with higher IoU will be given higher weighting. Formally we denote λ_c, λ_p and λ_d as the weights of center-aware consistency loss, probability-aware consistency loss and size-aware consistency respectively, $\{c^i, p^i, d^i\}$ as the center, score and area of i-th predicted box, we have:

$$\begin{aligned}\mathcal{L}_{cpd} \;&= \lambda_c \frac{\sum_i^{|B_t|} w_i \left\| c_t^i - c_a^i \right\|_2}{|B_t|} + \lambda_p \frac{\sum_i^{|B_t|} w_i \left\| p_t^i - p_a^i \right\|_2}{|B_t|} \\ &+ \lambda_d \frac{\sum_i^{|B_t|} w_i \left\| d_t^i - d_a^i \right\|_2}{|B_t|}\ , \end{aligned} \qquad (4)$$

where $w_i = IoU(b_t^i, b_a^i)$ equals to the IoU of pair $\{b_t^i, b_a^i\}$. With the IoU weights, we can increase the gradients for the prediction pairs with high IoU, which we consider them as highly confident pairs; and we reduce the gradients for pairs with low IoU, which we consider as being false aligned due to detection errors.

3.3 Proposal Consistency Loss

Two-stage detectors, such as the Faster R-CNN shown in Figure 2, usually exploit region proposal network (RPN) to predict the proposals. Unfortunately, the gradients from consistency loss built over the prediction can not be back-forward to the RPN branch due to undifferentiable gradients, which would otherwise lead to sub-optimal semi-supervised learning since RPN can not learn from unlabeled data. Especially in densely packed scenes with a large number of small objects that are hardly recognizable, the performance of the detectors depends highly on the proposal accuracy of RPN. To enable RPN to learn from unlabeled data, we design a proposal consistency loss that can force all the proposals from the student model to approach closely to the ones from the teacher model, that is:

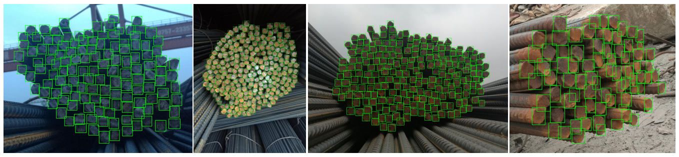

Figure 3: Examples and the box annotations in the RebarDSC dataset. The rebars are captured under different angles and illumination conditions.

$$\begin{aligned}\mathcal{L}_{rpn} \;&=\; \frac{\lambda_{r1}}{N} \sum_i^N \|dp_t^i\|_2 (\left\| dx_s^i - dx_t^i \right\|_2 + \left\| dy_s^i - dy_t^i \right\|_2 \\ &+ \left\| dw_s^i - dw_t^i \right\|_2 + \left\| dh_s^i - dh_t^i \right\|_2) \\ &+ \frac{\lambda_{r2}}{N} \sum_i^N \left\| dp_s^i - dp_t^i \right\|_2\ , \end{aligned} \qquad (5)$$

where λ_{r1} and λ_{r2} denote the weight of offset consistency and probability consistency, N denotes the number of proposals, $\{dp_t, dx_t, dw_t, dy_t, dh_t\}^N$ denotes the regression score and offsets of the teacher RPN, $\{dp_s, dx_s, dw_s, dy_s, dh_s\}^N$ denotes the regression score and offsets of the student RPN.

3.4 Optimization

In the semi-supervised learning, we optimize the student model with supervised loss from the labeled data and consistency loss from the unlabeled data at the same time. Formally, we denote the standard supervised loss from the detection framework, including classification loss and regression loss as $\mathcal{L}_{det}$, the overall losses can be formulated as:

$$\mathcal{L}_s \;= \frac{1}{|I_s \cup I_t|} \sum^{I_s \cup I_t} (\mathcal{L}_{iou} + \mathcal{L}_{cpd} + \mathcal{L}_{rpn}) + \frac{1}{|I_s|} \sum^{I_s} \mathcal{L}_{det}, \quad (6)$$

where $\mathcal{L}_s$ will be minimized with the optimization of the student model.

3.5 RebarDSC Dataset

Object detection in densely packed scenes has great potential to facilitate the job of product counting and improve quality control in the industry. To promote the exploration of the applications in densely packed scenes and to enable comprehensive performance evaluation, we build a new dataset of the industrial rebar collection scene, which is denoted as RebarDSC. To construct this dataset, (i) we collected 2,125 images from the top 3 rebar manufacturing companies in Asia, where the image are captured in the real production environment with various types of mobile devices, with resolution vary from 800×600 to 4600×3400; (ii) to satisfy the requirement of bundle counting, we capture all the raw images with a bundle of rebar as the image center. If two or more bundles are captured, we only considered the rebar within the bundle closing to the image center and ignored other bundles. (iii) we hired professional human workers from rebar manufacturing companies to annotate each image's bounding box, and finally, 350,348 annotations, 164.9 annotations per image are collected. Some samples in our dataset are shown in Figure 3.

Method	10%			20%			50%			100%		
	AP	AP$^{.75}$	AR^{300}	AP	AP$^{.75}$	AR^{300}	AP	AP$^{.75}$	AR^{300}	AP	AP$^{.75}$	AR^{300}
Backbone [Ren *et al.*, 2016]	48.2	51.4	55.6	49.7	53.9	53.7	52.0	57.8	59.2	53.8	60.5	60.7
CSD [Jeong *et al.*, 2019]	50.6	55.7	56.9	52.4	58.2	58.5	53.7	60.6	60.0	55.2	62.8	61.0
STAC [Sohn *et al.*, 2020b]	51.2	56.3	57.7	52.6	58.7	59.0	54.1	61.0	60.3	55.2	62.5	61.2
SESS [Zhao *et al.*, 2020]	50.8	55.7	57.1	52.1	57.7	58.4	52.9	58.7	59.9	52.8	59.1	60.4
Ours	**52.8**	**59.2**	**59.1**	**54.1**	**61.4**	**60.4**	**54.8**	**62.6**	**61.2**	**56.3**	**64.9**	**62.5**
Gain	**4.6**	**7.8**	**3.5**	**4.4**	**7.5**	**6.7**	**2.8**	**4.8**	**2.0**	**2.5**	**4.4**	**1.8**

Table 1: Quantitative results of semi-supervised object detection on the SKU-110K dataset.

Method	10%			20%			50%			100%		
	AP	AP$^{.75}$	AR^{300}	AP	AP$^{.75}$	AR^{300}	AP	AP$^{.75}$	AR^{300}	AP	AP$^{.75}$	AR^{300}
Backbone [Ren *et al.*, 2016]	57.3	67.7	63.9	59.4	71.2	66.4	61.1	74.6	67.6	62.8	76.9	68.8
CSD [Jeong *et al.*, 2019]	56.7	66.2	62.5	60.6	72.8	65.2	62.9	77.5	67.6	64.6	79.1	69.0
STAC [Sohn *et al.*, 2020b]	57.3	67.4	63.5	59.6	71.2	66.0	60.7	73.5	67.0	62.1	76.8	68.3
SESS [Zhao *et al.*, 2020]	57.2	67.1	63.1	59.3	71.3	65.2	61.7	75.3	66.9	62.9	77.6	68.6
Ours	**59.7**	**70.5**	**66.6**	**62.7**	**76.2**	**68.2**	**64.0**	**78.9**	**69.6**	**65.7**	**81.6**	**71.1**
Gain	**2.4**	**2.8**	**2.7**	**3.3**	**5.0**	**1.8**	**2.9**	**4.3**	**2.0**	**2.9**	**4.7**	**2.3**

Table 2: Quantitative results of semi-supervised object detection on the RebarDSC dataset.

4 Experimental Results

Implementation details. Our framework was implemented in PyTorch, using one NVIDIA GeForce 3090. We initialize Faster R-CNN with the parameters trained from COCO dataset [Lin *et al.*, 2014], and pre-train Faster R-CNN using all the available labeled samples with 12 epochs following the standard supervised learning. We then initialize the student and teacher networks with the pre-trained weights. With the initialized models, we train the student network on both the labeled and unlabeled data by minimizing the supervised loss and consistency losses with extra 12 epochs. The student network is trained by an SGD optimizer with momentum=0.9, weight decay=0.0001 and learning rate=0.0025. Each training batch contains three samples, consisting of one labeled sample and two unlabeled samples. All the input images are resized with their shorter side as 1200. The weights in the consistency loss functions are set as $\lambda_{iou} = 1, \lambda_c = 0.5, \lambda_p = 1, \lambda_d = 2, \lambda_{r1} = 2, \lambda_{r2} = 1$, which are chosen by cross-validations on the training set. Following the previous works [Laine and Aila, 2016; Tarvainen and Valpola, 2017], we also ramp up the coefficient of consistency loss and gradually increase the EMA decay factor λ_{ema} from 0 to 0.99 in all the epochs.

Datasets. We evaluate our framework on SKU-110K [Goldman *et al.*, 2019] and RebarDSC datasets. SKU-110K is a large public retail environment dataset collected from supermarket stores. It contains 8,233 images in the training set and 2,941 images in the test set. Each image in SKU-110K dataset averagely contains 149.4 annotations, which creates a challenge for detectors. For the RebarDSC dataset, we randomly select 1,000 images as the training set, and consider other 1,125 images as the test set.

Metrics. We adopt evaluation metrics similar to those used by COCO [Lin *et al.*, 2014], reporting the average precision AP at IoU=0.5:0.05:0.95, AP at IoU=0.75, and average recall AR^{300} at IoU=0.50:0.05:0.95, where 300 denotes the maximal number of objects.

4.1 Quantitative Comparison

Compared Methods. Since we are the first one to explore the semi-supervised learning for densely packed object detection, we compare the proposed method with the SSL object detection methods closest to ours: CSD [Jeong *et al.*, 2019], SESS [Zhao *et al.*, 2020] and STAC [Sohn *et al.*, 2020b]. CSD and STAC are designed for the SSL generic object detection, and the SESS is proposed to learn from 3D data. We try our best to implement these methods and carefully finetune the training parameters to obtain the best results. For STAC, we remove the data augmentations of geometric transformation and cutout which may destroy the densely packed objects; for SESS designed for 3D object detection, we modify the consistency loss from the 3D type into 2D. For all the compared methods, we replace their Backbone with Faster R-CNN+ResNet50+FPN, which is also adopted in our method for a fair comparison.

Table 1 shows the quantitative results of compared methods on the SKU-110K dataset. We mark the ratio of available labeled in the table; for example, 10% denotes that 10% of training data are used as labeled data, and the other training data are used as unlabeled data. Note that on the setting with 100% label data variable, we treat all the labeled data as unlabeled data and apply the proposed consistency losses over all the data. We also report the model trained with labeled data only and mark the model as “Backbone”. The results show that our method outperforms the other compared methods on all training data settings. Especially on the metric AP$^{.75}$, our method outperforms the backbone with 7.8%, 7.5%, 4.8% and 4.4%, indicating that our method can improve the localization accuracy of results. We argue that this is because our

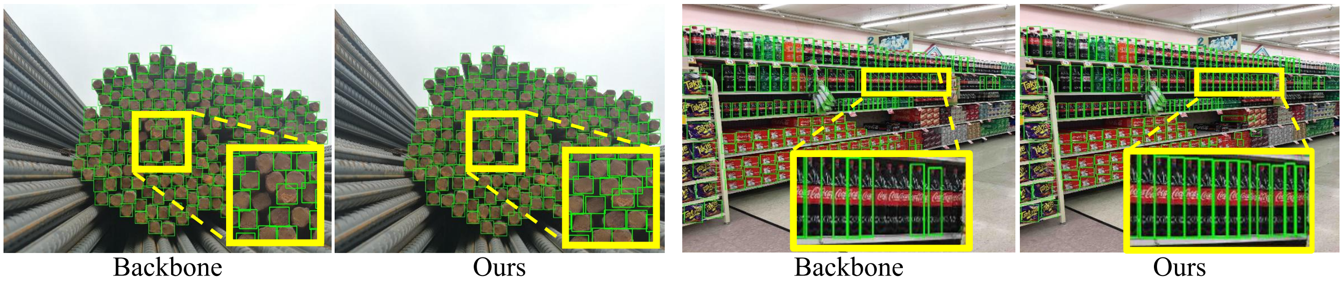

Figure 4: Visual comparison of our method and Backbone. Left: Rebar dataset; right: SKU-110K dataset.

Method	10%	20%	50%	100%
Backbone	48.2	49.7	52.0	53.8
IoU-aware Consist. Only	52.2	53.5	54.6	55.7
Proposal Consist. Only	51.9	53.3	54.7	55.5
Our Full	52.8	54.1	55.0	56.3

Table 3: Ablation study on the SKU-110K dataset. We report the AP in this table.

Detection Model	Backbone	Ours	Gain
Faster R-CNN+R50	49.7	54.1	4.4
Faster R-CNN+R101	49.9	54.5	4.6
RetinaNet+R50	47.4	49.5	2.1
RetinaNet+R101	47.9	51.0	3.1

Table 4: Evaluation of our method with different backbones Faster R-CNN and RetinaNet on the SKU-110K dataset (20%). We report the AP in this table.

proposed consistency loss can benefit the learning of densely object with diverse shapes in the market scenes.

Table 2 shows the quantitative results of compared methods on the RebarDSC dataset. Our method can significantly improve the detector over the backbone compared with other SSL methods. Especially on the metric 100%, our method can significantly outperform the backbone model, which shows that our methods can improve the detector performance in densely packed scenes with the help of the proposed consistency losses.

4.2 Visual Comparisons

To demonstrate the effect of our framework more intuitively, we visualize the detection results from our method and Backbone network on the SKU-110K dataset and RebarDSC dataset in Figure 4. Backbone here is trained with 10% labeled data available, and our method is trained with 10% labeled data and 90% unlabeled data. From the results, we can see that our method generally detects out more objects while the Backbone easily misses the objects with weak visual appearance. This proves that our method can let the detector learn more discriminative features from the unlabeled data.

4.3 Ablation Study

To analyze the effectiveness of IoU-aware consistency loss and proposal consistent loss in the proposed semi-supervised learning framework, we perform an ablation experiment on the SKU-110K dataset and report the results in Table 3. In this table, “IoU-aware Consistency Only” means that we implement our method without proposal consistency loss, and “Proposal Consistency Only” means that we implement our method without IoU-aware consistency loss.“Full” indicates that we implement our method with all the proposed losses. In this table, our method with either IoU-aware consistency losses or proposal consistency loss outperforms the backbone results, shows that both of the two proposed losses can effectively improve the model robustness with the learning of unlabeled data.

To validate that our method is general enough to be applied to different detectors, we evaluate our method with different backbones in Table 4. The results show that our method can be adapted to two-stage detector [Ren *et al.*, 2016] or one-stage detector [Lin *et al.*, 2017b].

5 Conclusion

In this paper, we present an approach to learn unlabeled data in densely packed scenes. Based on the consistency learning, we design the IoU-aware consistency loss to enforce the IoU consistency of prediction pairs, which can significantly improve the localization accuracy. We also observe that a large number of false aligned pairs existed in densely packed scenes; therefore, we reweight the prediction pairs with IoU so that the distraction from low-confident pairs can be eliminated. We further design the proposal consistency loss to encourage the consistency between the proposal. In this way, the region proposal network can be learned from unlabeled data. We also construct a new dataset RebarDSC to enrich the datasets in densely packed scenes. We test our method on two datasets, and the extensive results show that our method outperforms the other methods in densely packed scenes.

Acknowledgements

The work is supported by Key-Area Research and Development Program of Guangdong Province, China (2020B010165004, 2020B010166003, 2018B010107003), NSFC (61772206, U1611461, 61472145), Hong Kong Innovation and Technology Fund (MRP/015/18), Hong Kong Research Grants Council (PolyU 152006/19E, 152035/17E), and the Hong Kong Polytechnic University (PolyU 152009/18E).

References

[Berthelot *et al.*, 2019] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Proc. NeurIPS*, pages 5049–5059, 2019.

[Cai *et al.*, 2019] Qi Cai, Yingwei Pan, Chong-Wah Ngo, Xinmei Tian, Lingyu Duan, and Ting Yao. Exploring object relation in mean teacher for cross-domain detection. In *Proc. CVPR*, pages 11457–11466, 2019.

[Chen *et al.*, 2020] Zhiming Chen, Kean Chen, Weiyao Lin, John See, Hui Yu, Yan Ke, and Cong Yang. Piou loss: Towards accurate oriented object detection in complex environments. In *Proc. ECCV*, pages 195–211. Springer, 2020.

[Gao *et al.*, 2019] Jiyang Gao, Jiang Wang, Shengyang Dai, Li-Jia Li, and Ram Nevatia. Note-rcnn: Noise tolerant ensemble rcnn for semi-supervised object detection. In *Proc. ICCV*, pages 9508–9517, 2019.

[Goldman *et al.*, 2019] Eran Goldman, Roei Herzig, Aviv Eisenschtat, Jacob Goldberger, and Tal Hassner. Precise detection in densely packed scenes. In *Proc. CVPR*, pages 5227–5236, 2019.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, pages 770–778, 2016.

[Hong *et al.*, 2020] Fa-Ting Hong, Wei-Hong Li, and Wei-Shi Zheng. Learning to detect important people in unlabelled images for semi-supervised important people detection. In *Proc. CVPR*, pages 4146–4154, 2020.

[Hsieh *et al.*, 2017] Meng-Ru Hsieh, Yen-Liang Lin, and Winston H Hsu. Drone-based object counting by spatially regularized regional proposal network. In *Proc. ICCV*, pages 4145–4153, 2017.

[Jeong *et al.*, 2019] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. In *Proc. NeurIPS*, pages 10759–10768, 2019.

[Jiang *et al.*, 2018] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. In *Proc. ECCV*, pages 784–799, 2018.

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

[Lin *et al.*, 2014] 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 *Proc. ECCV*, pages 740–755. Springer, 2014.

[Lin *et al.*, 2017a] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proc. CVPR*, pages 2117–2125, 2017.

[Lin *et al.*, 2017b] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proc. ICCV*, pages 2980–2988, 2017.

[Pan *et al.*, 2020] Xingjia Pan, Yuqiang Ren, Kekai Sheng, Weiming Dong, Haolei Yuan, Xiaowei Guo, Chongyang Ma, and Changsheng Xu. Dynamic refinement network for oriented and densely packed object detection. In *Proc. CVPR*, pages 11207–11216, 2020.

[Ren *et al.*, 2016] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE TPAMI*, 39(6):1137–1149, 2016.

[Sohn *et al.*, 2020a] 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:2001.07685*, 2020.

[Sohn *et al.*, 2020b] 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:2005.04757*, 2020.

[Tarvainen and Valpola, 2017] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proc. NeurIPS*, pages 1195–1204, 2017.

[Wang *et al.*, 2020a] Dong Wang, Yuan Zhang, Kexin Zhang, and Liwei Wang. Focalmix: Semi-supervised learning for 3d medical image detection. In *Proc. CVPR*, pages 3951–3960, 2020.

[Wang *et al.*, 2020b] He Wang, Yezhen Cong, Or Litany, Yue Gao, and Leonidas J Guibas. 3dioumatch: Leveraging iou prediction for semi-supervised 3d object detection. *arXiv:2012.04355*, 2020.

[Wang *et al.*, 2020c] Jianfeng Wang, Xi Yin, Lijuan Wang, and Lei Zhang. Hashing-based non-maximum suppression for crowded object detection. *arXiv:2005.11426*, 2020.

[YuanQiang *et al.*, 2020] Cai YuanQiang, Dawei Du, Libo Zhang, Longyin Wen, Weiqiang Wang, Yanjun Wu, and Siwei Lyu. Guided attention network for object detection and counting on drones. In *Proc. Multimedia*, pages 709–717, 2020.

[Zhao *et al.*, 2020] Na Zhao, Tat-Seng Chua, and Gim Hee Lee. Sess: Self-ensembling semi-supervised 3d object detection. In *Proc. CVPR*, pages 11079–11087, 2020.