Group R-CNN for Weakly Semi-supervised Object Detection with Points

Shilong Zhang[1*], Zhuoran Yu[2*], Liyang Liu[3*], Xinjiang Wang[4], Aojun Zhou[4], Kai Chen [1,4]

[1]Shanghai AI Laboratory [2] Georgia Institute of Technology

[3] Tencent AI Platform Department, China [4] SenseTime Research

zhangshilong@pjlab.org.cn, zhuoranyu@gatech.edu, leonlyliu@tencent.com,

{wangxinjiang,zhouaojun,chenkai}@sensetime.com

Abstract

We study the problem of weakly semi-supervised object detection with points (WSSOD-P), where the training data is combined by a small set of fully annotated images with bounding boxes and a large set of weakly-labeled images with only a single point annotated for each instance. The core of this task is to train a point-to-box regressor on well-labeled images that can be used to predict credible bounding boxes for each point annotation. We challenge the prior belief that existing CNN-based detectors are not compatible with this task. Based on the classic R-CNN architecture, we propose an effective point-to-box regressor: Group R-CNN. Group R-CNN first uses ***instance-level proposal grouping*** *to generate a group of proposals for each point annotation and thus can obtain a high recall rate. To better distinguish different instances and improve precision, we propose instance-level proposal assignment to replace the vanilla assignment strategy adopted in original R-CNN methods. As naive instance-level assignment brings converging difficulty, we propose* ***instance-aware representation learning*** *which consists of instance-aware feature enhancement and instance-aware parameter generation to overcome this issue. Comprehensive experiments on the MS-COCO benchmark demonstrate the effectiveness of our method. Specifically, Group R-CNN significantly outperforms the prior method Point DETR by 3.9 mAP with 5% well-labeled images, which is the most challenging scenario. The source code can be found at https://github.com/jshilong/GroupRCNN.*

1. Introduction

Object detection has witnessed great improvements with the development of network architectures and dataset construction in the past few years. However, advanced object detectors usually require training on a large number of images with accurate bounding box annotations, which are very time-consuming and expensive to obtain. To relieve the burden of human labeling, previous studies propose weakly- [2, 11, 13, 21, 32] and semi-supervised methods [8, 18, 24, 31] for object detection, which adopts a small portion of well-annotated images alongside abundant weakly-annotated images and unlabeled images, respectively. Weakly semi-supervised object detection (WSSOD) [28] takes goods from both worlds. By replacing the unlabeled images in semi-supervised detection with weakly-labeled images, it achieves a good balance between labeling costs and model performance. One step further, WSSOD with point annotations (WSSOD-P) labels an instance with a single point so as to provide both category and location information with minimal labeling cost. The time cost of labeling points is comparable with that of providing image-level annotations [1, 4], which is substantially lower than bounding box annotations.

Prior work [4] proposes the following pipeline to leverage point annotations. **(1)** train a point-to-box regressor using images with only bounding box annotations. To simulate the process of translating points into bounding boxes, points are randomly sampled within bounding boxes of instances as point annotations. **(2)** After training, the regressor is used to transform point annotations on weakly-labeled images to pseudo bounding box annotations. **(3)** Finally, any object detector can be trained with both well-labeled images and pseudo-labeled images in a supervised fashion. The core of this task then boils down to designing an accurate point-to-box regressor.

Point DETR [4] claims that the CNN-based detector performs poorly as a point-to-box regressor, but we argue it does not hold true based on the following in-depth analysis. On one hand, we find that CNN models in prior work only generate a single proposal with the feature vector **located at the annotated point** (projected on the feature map). The quality of such a proposal can be quite low, resulting in a low recall rate since annotated points can deviate from object centers and are less informative for box regression. To

*Equal contribution

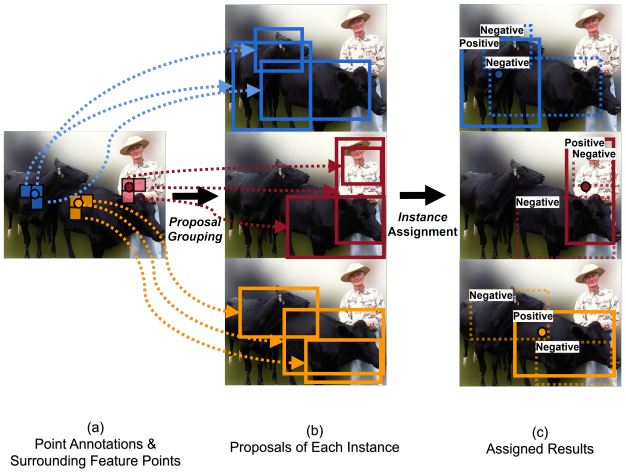

Figure 1. **An illustration of instance-level proposal grouping and instance-level proposal assignment**. Proposals in the same group (i.e. belonging to the same instance) are denoted by the same color. (a) to (b): **Instance-level grouping**: rather than just only the annotated point, proposals generated by all points that are close enough to the (projected) annotation point are collected to form a group for each instance. (b) to (c): **Instance-level assignment**: A proposal is assigned as positive if and only if the IoU with its *belonging* instance is above a pre-defined threshold, regardless of its IoU with any other instances.

improve the recall rate with arbitrarily annotated points, we propose instance-level proposal grouping (Figure 1, (a) to (b)) that generates a group of proposals rather than a single proposal for each instance. That said, points on feature maps that are **close to** the annotated point will be taken into consideration, and all proposals generated from these points form an instance-level group for a certain instance.

On the other hand, the original proposal assignment strategy assigns a proposal to be positive if its maximal IoU with **any ground truth boxes** surpasses a certain threshold. However, proposals in one group may be assigned to the ground-truth bounding box of another group when objects with the same category are close to each other (e.g. in the crowded scene). Such a scenario is very common in real-world datasets, for example, in MS-COCO over 50% images have such a scene. Training with such a strategy cannot ensure a high-quality proposal is produced for each point annotation. Therefore, we propose instance-level assignment that assigns proposals in a group only to its **belonging ground-truth box**. A proposal will be treated as negative if its IoU with the corresponding instance is low, regardless that its IoU with another instance may surpass the threshold (Figure 1, (b) to (c)).

Nevertheless, we find that the naive instance assignment causes converging difficulty and thus leads to inferior performance than the original assignment strategy. Since proposals from different groups share **the same feature map** and are also convolved with **the same fixed parameters**, positive proposals in one group and negative proposals in another group can have similar convolution outputs. However, they are assigned to completely opposite optimization targets, this contradiction can hinder model convergence. To tackle this problem, we propose instance-aware representation learning, consisting of instance-aware feature enhancement and instance-aware parameter generation. Specifically, we use point annotations to compute instance-aware relative coordinates, with which to construct instance-aware feature maps. To further distinguish features of proposals from different groups, we use both instance-aware features and the annotated points' category embeddings to generate instance-aware model parameters. By convolving the **instance-aware features** and the **instance-aware parameters**, we successfully mitigate the converging issue brought by naive instance assignment and obtain superior performance.

To highlight our key designs and the proposed general framework as a point-to-box regressor, we name our regressor Group R-CNN. Compared with the previous state-of-the-art method Point DETR which is based on the transformer architecture, our proposed Group R-CNN has the following advantages: (1) Group R-CNN can take advantage of the feature pyramid network [14] for multi-scale proposal generation, while Point DETR [4] can not naively adopt FPN. (2) Group R-CNN inherits the convergence superiority from CNNs, and thus converges better and faster than Point DETR, especially under the low well-labeled data regime [6] (See Appendix for details).

We conduct extensive experiments on MS-COCO dataset, using various percentages of labeled images to showcase the effectiveness of Group R-CNN. Group R-CNN outperforms the existing transformer-based method Point DETR by a large margin under different experimental protocols while only requiring half of the training budget. To be specific, our Group R-CNN outperforms Point DETR by 3.9 mAP when only 5% images are well-labeled by bounding boxes. Thus, Group R-CNN achieves both better performance and faster convergence over previous state-of-the-art.

2. Related Work

Supervised Object Detection. Supervised object detection has made remarkable progress over the past few years [3, 12, 15, 20, 26]. Based on the architecture design, these object detectors can be mainly categorized as two-stage detectors and single-stage detectors. Two-stage detectors [20] first generate a large set of object proposals and refine them in the second stage. In contrast, single-stage detectors [15, 17, 19] directly predicts bounding boxes with categories without refinement. In addition to designs of ob-

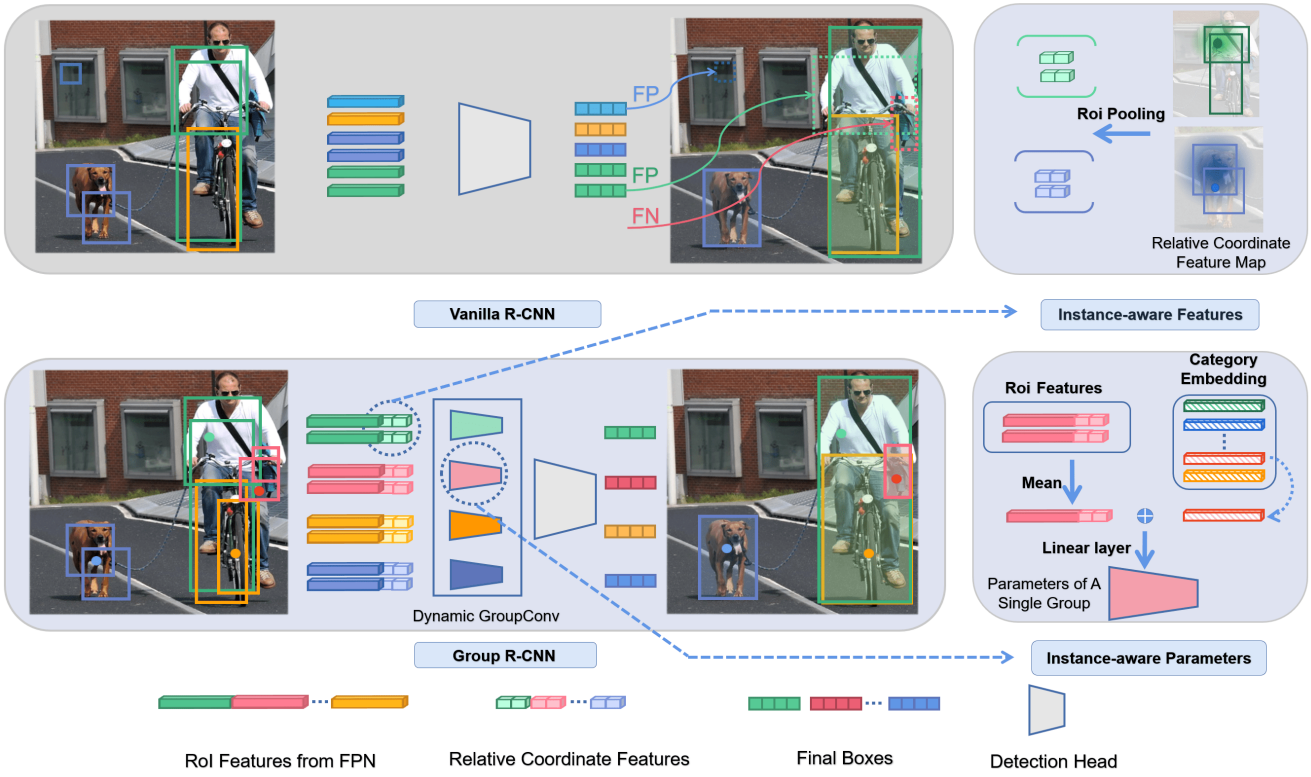

Figure 2. **The pipeline of Group R-CNN**. The Vanilla R-CNN (**the upper architecture**) cannot leverage the point annotation and easily produces false positive and false negative bounding boxes. In contrast, Group R-CNN (**the lower architecture**) achieves higher recall and precision by incorporating relative coordinates (top-right) and dynamic group convolution (bottom-right) for the point-to-box translation task. **Instance-aware Features (top-right)**: a separate feature map is constructed for each instance to encode the relative coordinates w.r.t. its point annotation, followed by the RoI-pooling procedure. **Instance-aware Parameters (bottom-right)**: the average RoI feature of each group is concatenated with the corresponding category embedding, acting as the input of a linear layer to generate parameters for the dynamic group convolution.

ject detectors, multi-scale object detection is another popular line of work. Feature Pyramid Network [14] produces multi-scale feature representation that can be used by both single-stage and two-stage object detectors

Semi- and Weakly-Supervised Object Detection. To reduce the annotation cost of producing large-scale well-labeled images for object detection, two popular settings were proposed. Semi-supervised object detection [18, 22, 31] replaces the weakly-labeled images with unlabeled images and leverages pseudo-labeling for better performance. The problem considered in this work is different from these two classic learning paradigm as the point annotation is added to weakly-labeled images. Weakly-supervised object detection [2, 9, 13, 21, 23, 27, 28] proposes to use a large set of weakly-labeled images (such as image-level annotations without bounding boxes) alongside a small set of well-labeled images. PCL [23] incorporates a concept of grouping and clusters proposals in an unsupervised manner to prevent from focusing on discriminative parts of instances. In contrast, Group R-CNN selects proposals around the human-labeled points to **improve the recall rate**, which is different from PCL in terms of motivation and implementation. Details of our method can be found in Section 3.

Weakly Semi-Supervised Object Detection with Points (WSSOD-P). Recently, Point DETR [4] proposes a new annotation format for weakly-supervised object detection, which adds point annotations as a new form of weak annotations. To leverage point annotations, Point DETR introduces a transformer-based point-to-box regressor to transform point annotations to bounding-box annotations so that an actual object detector can be trained. However, transformer-based models usually have convergence issues when training data is not sufficient [6], which is usually the case of WSSOD-P (See Appendix for details). In contrast, we design a convolution-based regressor for this task, which is considered challenging by prior work [4].

Dynamic Parameter Generation. Dynamic Parameter Generation has been used in various computer vision tasks

such as image classification [5,29,30] and instance segmentation [25]. The key idea behind this line of work is to dynamically adjust network parameters based on inputs. Our proposed dynamic group convolution is inspired by these existing works and we propose to generate convolution parameters for each instance-level proposal groups. To the best of our knowledge, this is the first time such an idea is implemented under WSSOD-P setting.

3. Group R-CNN

In this section, we first review the problem of weakly semi-supervised object detection with point annotations (WSSOD-P). Next, we present our novel framework Group R-CNN as a solution to this task.

3.1. Background

The problem of WSSOD-P proposes to train an object detector using a small portion of **well-labeled** images with instance-level annotations (e.g., bounding boxes and class labels) and an abundant **weakly-labeled** images with only one single point annotation for each instance. Compared with vanilla weakly-supervised object detection, the point annotation provides meaningful location information for instances without introducing much labeling costs. A common pipeline [4] to solve this problem is: (1) training a point-to-box regressor using well-labeled images with bounding box annotations (2) generating pseudo bounding boxes for images with point annotations, and (3) training an object detector with the combination of well-labeled and pseudo-labeled images. The core of this task is to design an effective point-to-box regressor that translates point annotations to credible pseudo bounding box annotations.

To design a better point-to-box regressor, we propose Group R-CNN, a CNN-based architecture for this point-to-box translation task. Group R-CNN inherits a multi-stage architecture as in Cascade R-CNN [3], which consists of a proposal generation stage and a proposal refinement stage. We introduce our novel designs of our architecture in detail in the next sections.

3.2. Instance-level Proposal Grouping

Previous attempts of using a CNN-based framework as point-to-box regressors generate a single proposal using the feature of the projected annotation point for an instance, which may lead to a inferior recall rate resulting from inaccurate point annotations. To boost the recall rate, we propose instance-level proposal grouping (Figure 3), which aggregates proposals generated by feature points close to a certain annotation point to form a group. This strategy is based on the insight that points on feature maps corresponding to high-quality proposals are often close to the projection of annotation points on feature maps. Specifically, we collect k feature points around the projected annotation point on each level of the feature pyramid. Proposals produced by these km points then form a group for a given instance, where m is the number of levels in FPN. Each group finally produces nkm proposals where n is the number of anchors with different scales and aspect ratios at each point in RPN. The proposed instance-level proposal grouping collects information from neighboring points around the annotated point and thus can improve recall rate and is more robust to the inaccurate annotation points.

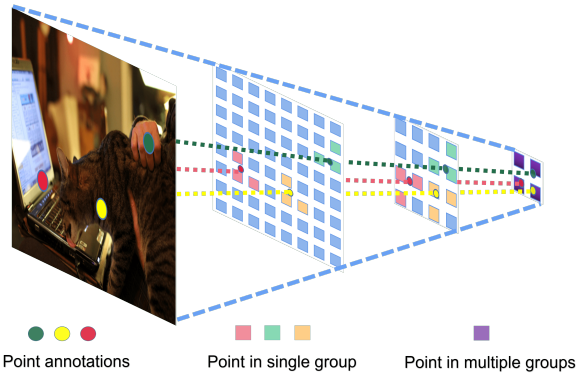

Figure 3. **Instance-level Proposal Grouping**. Proposals generated by points that are adjacent to projected annotation points form a single group. Such points can present in multiple groups.

3.3. Instance-aware Representation Learning

Instance-level Proposal Assignment. To achieve high recall while retraining high precision, the regressor is supposed to produce a single accurate bounding box for each point annotation. Therefore, we only output one proposal with the highest prediction score in a single group. However, the vanilla designs in classic R-CNN architectures have difficulty achieving such a goal especially when **objects with the same category are adjacent to each other** (crowded scene problem). First, in training, with the vanilla assignment strategy, proposals are regarded as positive if their maximal IoU with **any ground truth boxes** is greater than a pre-defined threshold and are considered as negatives otherwise. However, when objects of the same class are close to each other, proposals within one group may be matched to the ground-truth bounding box of another instance and eventually harm the precision. To tackle this problem, we propose instance-level assignment so that proposals of a certain group can only be assigned to the corresponding ground-truth bounding box of this group. Proposals are treated as negative if the IoU between them and the corresponding instance is below the threshold even though they might reach the IoU threshold with ground-truth bounding boxes of another group.

However, the crowded scene problem cannot be solved by naive instance-level assignment (as in Table 1). Since

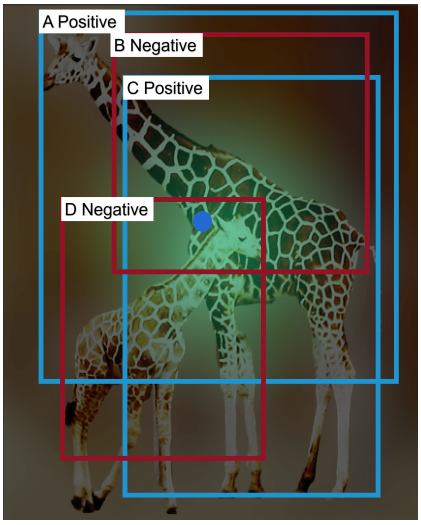

Figure 4. **An illustration of overlapped instances with the same category**. All four proposals are in the group of the taller giraffe. With the instance-level assignment, only proposal A and C are treated as positive. However, proposal D exposes a high-level similarity with positives of the group of the baby giraffe. **Because proposals in different groups share the feature pyramid and shared network parameters are used to process these features**, the predictions of positive proposals in the group of the baby giraffe and negative proposals in the group of tall giraffes are similar to each other, which causes confusion in model training.

proposals in different groups share the same feature pyramid and the network also uses the shared parameters of R-CNN heads to process these features, positive proposals in one group and negative proposals in another group can have similar processed results, but are assigned to completely opposite optimization targets (as in Figure 4), which causes difficulty in model training. To address this difficulty, we propose our solutions from two different aspects: **instance-aware feature enhancement** through relative coordinates encoding and **instance-aware parameters** for a dynamic group convolution.

Table 1. Failure of the naive instance-level assignment. Both methods are equipped **with instance-grouping** and output the best proposal of highest prediction score after NMS for each group.

	mAP	AP@50	AP@75
Vanilla Assign	36.6	61.5	37.7
Instance Assign	34.2	60.2	34.9

Instance-aware Feature Enhancement. To achieve instance-aware feature enhancement, we introduce relative coordinates encoding. Specifically, we leverage a prior that proposals closer to a point annotation should have a higher chance to be assigned to this instance. Specifically, a feature map $f \in \mathcal{R}^{H\times W\times 2}$ is constructed for each instance to encode the relative coordinate offset to its point annotation $f_{ij} = [\Delta x_{ij}, \Delta y_{ij}]^T$, where Δx_{ij} and Δy_{ij} denote the coordinate offset of each pixel at index i, j to the corresponding point annotation. The feature f also follows the same RoI pooling procedure for each proposal and then gets concatenated with the normal RoI-pooled feature, as shown in Figure 2. Due to the different coordinates of each group, instance-aware RoI features for different groups are generated even they have similar appearances, which is not uncommon in crowded scenes.

Instance-aware Parameter Generation. We introduce instance-aware model parameters so that feature representations of proposals in different groups can be further diversified. Although a single proposal may have a larger IoU with the ground-truth bounding box of another group, for most proposals in a group, the maximum IoU is still achieved with the corresponding ground-truth instance. Thus, we calculate the mean of all RoI features in a group as the representation of the instance. More specifically, we first collect RoI features of all proposals in a group and apply spatial average pooling on them to produce vectorized features. Then, we compute the mean over these features to form the instance representation. Furthermore, to better leverage the category information of the point annotation, we introduce a class embedding matrix of shape $C \times 256$ where C is the number of classes. This embedding matrix is optimized together with other model parameters. The embedding vector of the instance class and the constructed instance representation are concatenated, with which we produce instance-aware model parameters. Then we convolve the dynamically generated parameters with the above instance-aware proposal features to produce the classification score in the context of the instance-level assignment. This process is illustrated in the bottom-right component in Figure 2 and summarized in Equation (1) where $0 \le i \le N-1$ and $0 \le j \le G-1$. N and G are the number of instances and size of groups respectively. f_j^i is the instance-aware feature of j-th proposal in i-th instance group and $f_j'^i$ is the corresponding output feature of dynamic group convolution. P_i is the generated parameter of the i-th group and $\mathbf{C}_i$ is the category embedding of the group i. $F(\cdot)$ stands for a linear projection with learnable parameters. The dynamic group convolution is a **high-order transformation** as the parameters are generated from averaged proposal features, which makes features from different groups more discriminative.

$$
\begin{aligned}
f_j'^i &= f_j^i \otimes P_i \\
P_i &= F(\frac{\sum_j f_j^i}{G}, \mathbf{C}_i)
\end{aligned} \tag{1}
$$

The dynamic convolutions can be efficiently implemented as a group convolution where the number of groups is equal to the number of instances in the image. Through

convolving instance-aware features with instance-aware parameters, we effectively increase the discrimination among instances. The generated convolution layer is used to process the RoI features of proposals in a certain group. The classification and regression heads are attached after our Dynamic Group Convolution to predict final class predictions and regression offsets.

4. Experiments

Dataset. We evaluate the performance of Group R-CNN on MS-COCO [16] dataset. MS-COCO contains 118k training images with bounding box annotations and 5k images for validation. We randomly sample different percentages of images from the training set with fixed random seed 0 as our well-labeled set and use the rest as weakly-labeled sets with point annotation. To train our point-to-box regressor with well-labeled images, we randomly sample points within bounding boxes of instances at each iteration and use bounding boxes as the optimization target. After training, we run the inference process of the regressor on weakly-labeled images with point annotations. Following prior work [4], we synthesize the point annotations for weakly-labeled images by randomly sampling a point from instance masks of objects for one-shot. We also evaluate our method on Pascal VOC [7] and the results are in Appendix.

Architecture Details. Group R-CNN is built on top of Cascade R-CNN [3] and serves as the point-to-box regressor in WSSOD-P tasks. To better leverage the category label of point annotations, we adopt class-aware detector RetinaNet [15] as our region proposal network (RPN). Following the original RetinaNet, we use feature pyramid levels from P_3 to P_7 to extract region proposals. Notice that RetinaNet makes class-level predictions when extracting region proposals whereas the original RPN only produces objectness scores as in Faster R-CNN [20]. For RoI pooling in the second stage, we drop P_7 and only use P_3 to P_6 as candidates. For proposal refinement, a normal Cascade R-CNN head with three stages is used. Each stage has two shared FC layers and two separated FC layers to do the regression and classification, respectively. Unless otherwise specified, the model hyper-parameters are set as default [3]. RPN runs non-maximum suppression (NMS) with an IoU threshold of 0.7 on proposals from all groups with the same category based on the class prediction score and only 50 proposals are kept in each group. As for the final prediction results, we only select the proposal with the highest precision score in each group after NMS. In our dynamic group convolution, we always set the generated kernel size as 1 for simplicity.

Implementation Details. For a fair comparison, we follow the training setting of Point DETR [4] as much as possible. To train our Group R-CNN, we use 8 RTX 2080Ti GPUs with 2 images per GPU. Our data augmentation policy is also exactly the same as Point DETR. Though our proposed Group R-CNN has a different overall architecture with Point DETR [4], we use ResNet-50 [10] as our backbone, which is consistent with Point DETR. We use the learning rate of 0.02. We use SGD as our optimizer with a momentum of 0.9 and weight decay of $1e^{-4}$. The only difference in the training setting is the training schedule. Point DETR trains the regressor with 108 epochs and applies learning rate decay at the 72-th and 96-th epoch. **Our training schedule is substantially shorter than Point DETR**, we only train Group R-CNN by 50 epochs and decay the learning rate at the 30-th and 40-th epoch.

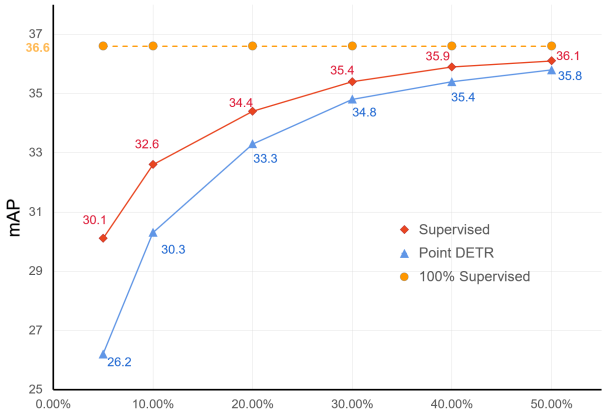

Figure 5. **Comparison with Point DETR.** Our method consistently outperforms Point DETR especially with limited well-labeled images. When using 50% well-labeled images, the FCOS model trained with well-labeled images and pseudo-labeled images produced by our method closely matches the supervised baseline trained with 100% well-labeled images.

4.1. Comparison with Point DETR

We also follow the evaluation setting of Point DETR where we train our point-to-box regressor Group R-CNN with well-labeled images and inference on weakly-labeled images. As a final step, we train an FCOS detector with both well-labeled images and weakly-labeled images, which is also consistent with Point DETR. The training setting of FCOS follows the standard 1x training schedule with exactly the same hyper-parameters as in the standard supervised training setting. We report the performance of the FCOS detector in Figure 5, which shows that Group R-CNN outperforms Point DETR with all different well-label fractions, especially when the well-labeled images are limited. We find that the fewer images are well labeled, the more Group R-CNN outperforms Point DETR. Specifically, in the most challenging scenario when only 5% and 10% images are labeled with bounding boxes, we achieve an improvement of 3.9 mAP and 2.3 mAP, respectively. Moreover, Group R-CNN only requires 50% training epochs of Point DETR to achieve such a large improvement. When

training with 50% of well-labeled images, FCOS trained with pseudo-labeled images produced by Group R-CNN achieves comparable performance as the model trained with 100% well-labeled images. It shows that we can largely close the gap between weakly semi-supervised detection and supervised detection with only point annotations.

4.2. Ablation Study

To illustrate the effectiveness of our proposed components, we conduct extensive ablation studies. In all of our ablation studies, Group R-CNN is trained with 10% well-labeled images. To eliminate interference factors and demonstrate the effectiveness of our proposed components on the point-to-box regressor, we report test results on the COCO validation set with fixed point annotation. In other words, we only evaluate the performance of the point-to-box regressor instead of training an object detector with produced pseudo-labeled images. We train our Group R-CNN only with 24 epochs in this section because it is already sufficient to show the effectiveness of our design.

Table 2. Impacts of sizes of different groups. Results are computed at IoU=0.5.

	AR	AR_s	AR_m	AR_l
No Grouping	80.4	61.1	86.8	95
$k = 1$	82.7	75.0	86.0	91.6
$k = 3$	**90.6**	85.1	93.7	96.9
$k = 6$	**92.5**	86.1	95.8	98.9
$k = 9$	**92.3**	84.8	95.5	99.2

Size of Instance Group. Recall that we select k points around the projected annotation point on each level of the feature map and collect the output proposals as a group. We conduct experiments with different k values to study how the size of instance groups impacts the recall of our point-to-box regression. For the baseline with no proposal grouping, we run NMS on all proposals with IoU=0.7 and select the top-1000 proposals with the highest classification score. For proposal grouping, we keep 50 proposals for each group after NMS. On average, each image from COCO contains roughly 7.27 objects. Thus, our method produces $7.27 \times 50 = 363.5$ proposals as expectation. As shown in Table 2, the performance of our RPN is significantly improved when instance grouping is adopted. With proposal grouping, although the model only produces around 30% proposals compared with that of the model without instance grouping, the AR is improved by more than 10%. Moreover, the performance is further boosted when k increases from 1 to 3. Although a larger k yields better results, $k > 3$ does not improve the recall by a large margin. Therefore, for a better performance-complexity trade-off, we choose $k = 3$ as the default choice in Group R-CNN.

Table 3. Ablation Study: Relative Coordinates

	mAP	AP@50	AP@75
w/o relative coordinates	35.7	61.0	37.0
w/ relative coordinates	37.1	64.1	38.1

Table 4. Impact of additional projection layers.

#proj. layers	mAP	AP@50	AP@75
0	34.2	60.2	34.9
1	**35.7**	61.0	37.0
2	35.8	60.9	36.9
3	35.7	60.6	36.7

Instance-Level Assignment. Now we compare our instance-level proposal assignment with the vanilla proposal assignment. For the vanilla assignment strategy, we assign the proposals which have a maximal IoU with any ground-truth bounding boxes larger than predefined thresholds to be positive. The IoU thresholds are set as 0.5, 0.6, and 0.7 for the three stages, respectively, following the default setting as Cascade R-CNN [3]. NMS with an IoU threshold of 0.5 is applied at the last stage and only the proposal with the highest class score for each group is kept. For the instance-level assignment, we use the same IoU thresholds as the vanilla counterpart.

However, we find that directly applying instance-level assignment causes convergence difficulty. Although the R-CNN refinement stages are trained with the instance-level assignment. RPN is still trained by the vanilla assignment. Then RPN and R-CNN will have conflict requirements for the FPN. Therefore, we first detach FPN from the R-CNN heads to block the back-propagated gradients. Then we add additional projection layers on FPN to produce a separate FPN designated for R-CNN.

We compare the effect of using different number of projection layers and the results are shown in Table 4. We show the effectiveness of additional projection convolution layers to process features with the detaching strategy. Using one additional projection layer lifts the mAP by 1.5 whereas more projection layers do not result in further improvements. Therefore, we simply use one additional projection layer so that the additional computation overheads are minimal.

Effectiveness of Instance-aware Feature Enhancement. As shown in Table 3, the mAP is improved from 35.7 to 37.1 when using relative coordinates. As discussed, although proposals from different instance groups may share similar appearances, enhancing the proposal feature with the relative coordinates endow the proposal with the ability to distinguish instances, given the prior that (ground truth) objects are located at different positions in the image.

Effectiveness of Instance-aware Parameter Generation. Table 5 shows the results when generating convolution

Table 5. Impact of Category Embedding (CE) and RoI Features (RoI-Feat.) on dynamic group convolution. In the baseline with 37.1 AP, all RoIs are not refined by dynamic convolution

CE	RoI-Feat.	mAP	AP@50	AP@75
		37.1	64.1	38.1
✓		38.5	64.7	39.9
	✓	38.2	64.5	39.7
✓	✓	39.2	65.7	41.0

Table 6. Comparison between **vanilla R-CNN with instance-level grouping** and **Group R-CNN** in crowded and non-crowded scenes, "vanilla" stands for the Vanilla R-CNN and "Group" denotes our proposed Group R-CNN.

dataset	Method	mAP	AP@50	AP@75
crowded 54k images	Vanilla	32.2	56.9	32.4
	Group	35.4	62.5	35.8
	Δ	+3.2	+5.6	+3.4
non-crowded 52k images	Vanilla	46.9	73.9	50.7
	Group	48.5	75.0	52.7
	Δ	+1.6	+1.1	+2.0

parameters with RoI features, category embeddings, and the concatenation of both, respectively. Clearly, generating parameters with both RoI features and category embeddings substantially outperforms using only a single component. Moreover, using a regular convolution layer in replacement of our dynamic group convolution only achieves 38.0 mAP (not included in the table), which is worse than dynamic group generation no matter what features are used in parameter generation. Therefore, our dynamic group convolution is a crucial design for Group R-CNN. We always set kernel size to 1 in all experiments since using larger kernel sizes such as 3 just gives marginal improvement (39.5 mAP for kernel size equals 3).

Compare with Vanilla Assignment. In real-world datasets, it is very common that objects of the same class have overlaps. For example, more than 50% images from MS-COCO contain crowded scenes. We compare the performance of Group R-CNN and vanilla R-CNN with instance-grouping under crowded scenes and non-crowded scenes. The models are evaluated on 90% weakly-labeled images. Table 6 shows that Group R-CNN(with our proposed instance-level assignment, instance-aware feature enhancement and instance-aware parameter generation) significantly outperforms the baseline by 3.2 mAP and 1.6 mAP respectively. It is obvious that Group R-CNN performs particularly better in crowded scenes, which well-supports the motivation of our method.

4.3. Visualization Results

We provide qualitative analysis on the validation dataset and visualize the detection results of our point-to-box regressor Group R-CNN compared with vanilla R-CNN. As shown in Figure 6, each bounding box is generated with one point annotation (indicated by the same color). Vanilla R-CNN produces a number of bounding boxes that are largely overlapped with other instances resulting in a low recall rate. In contrast, Group R-CNN successfully produces bounding boxes for most point annotations even in crowded scenes with objects of the same category.

Figure 6. **Visualization of predicted pseudo bounding boxes by Group R-CNN (ours) and vanilla R-CNN.** Group R-CNN achieves high precision and recall in the point-to-box regression task ensuring one accurate bounding box is produced for each point annotation. In contrast, vanilla R-CNN fails to capture lots of instances even with point annotation.

5. Acknowledgement

This project is supported by the Shanghai Committee of Science and Technology, China (Grant No. 20DZ1100800).

6. Conclusion

We propose Group R-CNN, a CNN-based point-to-box regressor for weakly semi-supervised object detection task. Group R-CNN leverages instance-level proposal grouping and instance-level representation learning(through instance-aware feature enhancement and instance-aware parameter generation) to improve the recall and precision. With these novel designs, Group R-CNN significantly outperforms existing transformer-based regressor by a large margin especially when well-labeled images are limited.

References

[1] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *European conference on computer vision*, pages 549–565. Springer, 2016. 1

[2] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2846–2854, 2016. 1, 3

[3] 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, 4, 6, 7

[4] Liangyu Chen, Tong Yang, Xiangyu Zhang, Wei Zhang, and Jian Sun. Points as queries: Weakly semi-supervised object detection by points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8823–8832, 2021. 1, 2, 3, 4, 6

[5] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11030–11039, 2020. 4

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3

[7] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 6

[8] Jiyang Gao, Jiang Wang, Shengyang Dai, Li-Jia Li, and Ram Nevatia. Note-rcnn: Noise tolerant ensemble rcnn for semi-supervised object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9508–9517, 2019. 1

[9] Wei Gao, Fang Wan, Xingjia Pan, Zhiliang Peng, Qi Tian, Zhenjun Han, Bolei Zhou, and Qixiang Ye. Ts-cam: Token semantic coupled attention map for weakly supervised object localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2886–2895, 2021. 3

[10] 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. 6

[11] Zequn Jie, Yunchao Wei, Xiaojie Jin, Jiashi Feng, and Wei Liu. Deep self-taught learning for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1377–1385, 2017. 1

[12] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750, 2018. 2

[13] Xiaoyan Li, Meina Kan, Shiguang Shan, and Xilin Chen. Weakly supervised object detection with segmentation collaboration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9735–9744, 2019. 1, 3

[14] 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. 2, 3

[15] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2, 6

[16] 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. 6

[17] 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. 2

[18] 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, 3

[19] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017. 2

[20] 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. 2, 6

[21] Miaojing Shi, Holger Caesar, and Vittorio Ferrari. Weakly supervised object localization using things and stuff transfer. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3381–3390, 2017. 1, 3

[22] 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. 3

[23] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan Yuille. Pcl: Proposal cluster learning for weakly supervised object detection. *IEEE transactions on pattern analysis and machine intelligence*, 42(1):176–191, 2018. 3

[24] Yuxing Tang, Josiah Wang, Boyang Gao, Emmanuel Dellandréa, Robert Gaizauskas, and Liming Chen. Large scale semi-supervised object detection using visual and semantic knowledge transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2119–2128, 2016. 1

[25] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 282–298. Springer, 2020. 4

[26] 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. 2

[27] Fang Wan, Chang Liu, Wei Ke, Xiangyang Ji, Jianbin Jiao, and Qixiang Ye. C-mil: Continuation multiple instance learning for weakly supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2199–2208, 2019. 3

[28] Ziang Yan, Jian Liang, Weishen Pan, Jin Li, and Changshui Zhang. Weakly-and semi-supervised object detection with expectation-maximization algorithm. *arXiv preprint arXiv:1702.08740*, 2017. 1, 3

[29] Brandon Yang, Gabriel Bender, Quoc V Le, and Jiquan Ngiam. Condconv: Conditionally parameterized convolutions for efficient inference. *arXiv preprint arXiv:1904.04971*, 2019. 4

[30] Yikang Zhang, Jian Zhang, Qiang Wang, and Zhao Zhong. Dynet: Dynamic convolution for accelerating convolutional neural networks. *arXiv preprint arXiv:2004.10694*, 2020. 4

[31] 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, 3

[32] Yi Zhu, Yanzhao Zhou, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Soft proposal networks for weakly supervised object localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1841–1850, 2017. 1