Group R-CNN for Weakly Semi-supervised Object Detection with Points (Supplementary Material)

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

In this supplementary material, we ask the following questions. Then we give answers to the above questions, one section for each question.

- Why we choose to develop a CNN-based model rather than a transformer-based one?
- How does vanilla assignment (instead of instance-level assignment) work with the proposed instance-aware representation learning?
- How can Group R-CNN be improved with weakly-labeled images (only with point annotations)?
- To what extent does Group R-CNN outperform semi-supervised object detection methods?
- Can Group R-CNN generalize well on other benchmarks like VOC?
- What are the limitation and negative social impacts of Group R-CNN?

1. Motivation: Inferior Performance of DETR When Data Lacks

One of our motivations to develop a CNN-based model is that transformer-based models cannot generalize well when trained with insufficient data. We provide experimental results to support this argument. Specifically, we train two representative CNN-based object detectors (Faster R-CNN [6] and RetinaNet [5]) alongside a transformer-based detector (DETR [2]) with various percentages of images (labeled with bounding boxes) in a supervised fashion. They have similar performance when the entire COCO dataset is used. However, CNN-based detectors, especially the two-stage detector, are significantly better than DETR when training data is limited (see Figure 1). The gap between CNN-based detectors and the transformer-based detector becomes larger when the number of labeled images decreases. It implies that the transformer-based model performs poorly in the case of data scarcity, but our pipeline involves training a point-to-box regressor with **limited data**, so it motivates us to develop a CNN-based regressor.

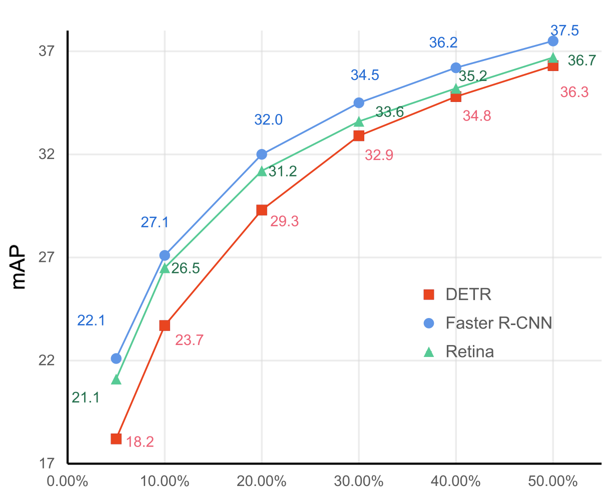

Figure 1. **Convergence Analysis of Different Detectors**

2. Vanilla Assignment with Instance-aware Representation Learning

In this work, we propose instance-aware feature enhancement and instance-aware parameter generation to comply with instance-level proposal assignment. We show that vanilla assignment is ineffective even with the above two strategies, demonstrating the necessity of instance-level assignment. We report the results of combining vanilla assignment with instance-aware feature enhancement and

[*]Equal contribution

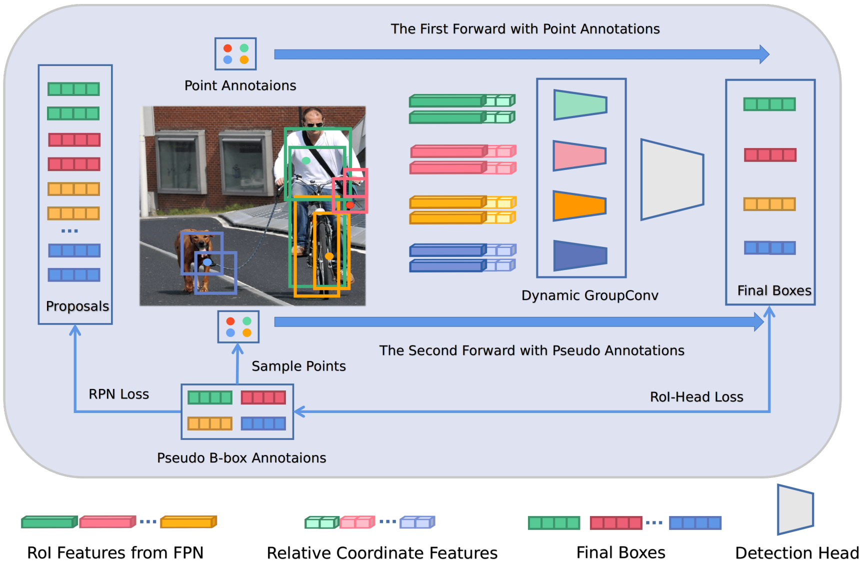

Figure 2. **The Pseudo-Labelling pipeline of Group R-CNN**. At each iteration, a weakly-labeled image is first forward to the model to obtain pseudo bounding box annotations. Next, a point is randomly sampled within each pseudo bounding box as pseudo point annotation. Finally, the model is trained with pseudo point annotations and pseudo annotations in the same fashion as training with well-labeled images.

instance-aware parameter generation. Specifically, we evaluate the performance of the **baseline** (using instance grouping and vanilla assignment) when additionally incorporating: (1) detaching FPN [4], (2) additional projection convolution (PC, also in Table 4, Section 4.2 of the main submission), (3) relative coordinates (RC) and (4) dynamic group convolution (DGC). For all experiments in this section, we exclude the default RoI sampling procedure to keep the size of instance group fixed to facilitate parallelization, so there is a slight performance drop (from 36.6 to 36.2) for Cascade R-CNN [1] when exploiting instance grouping.

Table 1 shows the results of adding each component to the baseline. It can be seen that most of our designs only bring marginal or no improvement over the baseline, and the detaching strategy even hurts the performance. When replacing instance-level assignment in Group R-CNN with the vanilla assignment and removing detaching (the last row in Table 1), the model only achieves 37.2 mAP, which is significantly lower than Group R-CNN with the instance-level assignment (39.2 mAP). Hence, the instance-level assignment is an essential building block of Group R-CNN.

3. Improving Group R-CNN with Weakly-Labeled Images

To compare fairly with Point DETR [3], we follow their setting and train our point-to-box regressor Group R-CNN with only well-labeled images. However, our proposed framework is general and thus it is also possible to exploit weakly-labeled images during training, similar to semi-supervised learning (pseudo-labeling). The pipeline is shown in Figure 2. At each iteration, for the weakly-labeled images, we first generate pseudo-bounding boxes with the Group R-CNN. Then, the generated boxes play the same role as those in well-labeled images. That said, we randomly sample a point within the pseudo bounding box to be the pseudo point annotation. Now the model can be trained with both weakly-labeled (with pseudo points as inputs and pseudo boxes as targets) and well-labeled images (with sampled points as inputs and human-labeled boxes as targets). We set the ratio between well-labeled images and weakly-labeled ones to 1:1 and the losses from weakly-labeled images are weighted by 0.5. We train Group R-CNN with the conventional multi-scale training for 50 epochs. Notice that we do not include any advanced strategies in the latest semi-supervised learning literature such as exponential moving average [**?**] and strong augmentation [**?**].

As shown in Table 2, training with both well-labeled and weakly-labeled images achieves a 2.6 mAP improvement. Even though our pseudo-labeling pipeline does not include any advanced designs in semi-supervised methods, the results already show that such a pipeline is plausible and could yield better performance than our point-to-box regressor trained without using the weakly-labeled images (with **only** point annotations).

4. Comparing with Vanilla Semi-Supervised Learning

In this work, we study weakly semi-supervised learning with point annotations. The key difference from vanilla

Table 1. Using Vanilla Assignment Strategy in Group R-CNN

	mAP	AP@50	AP@75
Casecade R-CNN + Instance Grouping	36.2	60.9	37.6
w/ deteaching	34.5	59.4	35.4
w/ 1 projection conv	36.6	60.6	38.4
w/ relative coordinates	36.2	60.7	37.7
w/ 1 dynamic group convolution	36.6	61.5	38.4
Group R-CNN w/ vanilla assign	37.2	61.6	38.8
Group R-CNN	**39.2**	**65.7**	**41.0**

Table 2. Training Group R-CNN in a Semi-Supervised Fashion

	mAP	AP@50	AP@75
w/o pseudo-labelling	39.5	66.5	41.1
w/ pseudo-labelling	42.4	69.0	44.7

semi-supervised learning is that instead of using **unlabeled** images without any form of annotations, we use **weakly-labeled** images with point annotations. We compare the performance of using weakly-labeled images and unlabeled images under STAC [7], a semi-supervised object detection pipeline. STAC first trains a teacher model with only well-labeled images and produces fixed pseudo bounding boxes for unlabeled images. Then the pseudo bounding boxes are used to train a student model. We replace the pseudo bounding box produced by STAC with the ones generated by Group R-CNN to demonstrate the advantage of point annotations. To keep the fairness of comparison, we adopt the same strategy and hyperparameters (including augmentation, training steps, batch size et al.) as STAC when training the student model using the generated offline pseudo bounding boxes. We evaluate the performance with 10% well-labeled images from MS-COCO.

Using pseudo bounding boxes produced by Group R-CNN improves the student mAP by 5.4 mAP (from 28.7 to 34.1), which shows that the quality of pseudo bounding boxes produced by point annotations is significantly better than those generated directly from unlabeled images. The experimental results validate the efficacy of point annotations under the semi-supervised setting.

5. Experiments on VOC datasets

We achieve 76.2 AP on VOC07+12 with 50 % point annotation and 50% box annotation. For reference, the AP of only 50% boxes and 100% boxes is 73.5% and 77.4%, respectively

6. Limitation and Social Impact

In this work, we follow the pipeline of Point DETR to tackle the problem of weakly semi-supervised object detection with points. Specifically, the pipeline involves: training a point-to-box regressor on well-labeled images, generating pseudo bounding boxes on weakly-labeled images, and training an object detector with the combination of well-labeled images and weakly-labeled images. The focus of this work is to develop a more accurate point-to-box regressor. However, when training such a regressor, only the well-labeled images are used. It is also possible to incorporate weakly-labeled images in the regressor training with a self-training fashion.

The potential social impact of this work inherits from object detection. Annotation costs are significantly reduced with weakly point annotations but with our method, it is still possible to train a promising object detector with significantly lower labeling costs. Consequently, undesired applications of object detection systems such as surveillance may be more accessible. Note that any advances in object detection and low-label learning paradigm could result in similar social impacts.

References

[1] 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

[2] 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

[3] 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. 2

[4] 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

[5] 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. 1

[6] 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

[7] 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