Supplementary Material for "Data-Uncertainty Guided Multi-Phase Learning for Semi-Supervised Object Detection"

1. Visualization of Label Noise Overfitting Problem

We select some images to further explain the label noise overfitting problem. The examples are in Fig. 1. In the first case, the supervised model mislabels the aeroplane as a car, and we notice that the semi-supervised model trained with this pseudo label detects one more car that does not exist. In the second case, the supervised model thinks a part of the iron mesh is a person. This may be an accident mistake. However, in the results from the semi-supervised model, more parts of the iron mesh are detected as persons. The same thing also happens in other cases. The noisy pseudo labels generated by the FS model are regarded as the correct ones by the SSOD model. The SSOD then overfits to these noisy labels, a nature of deep learning neural networks. In the test stage, the SSOD model just tries to intimate the noisy labels used for training, thus generates more similar mistakes. These examples validate the label noise overfitting problem.

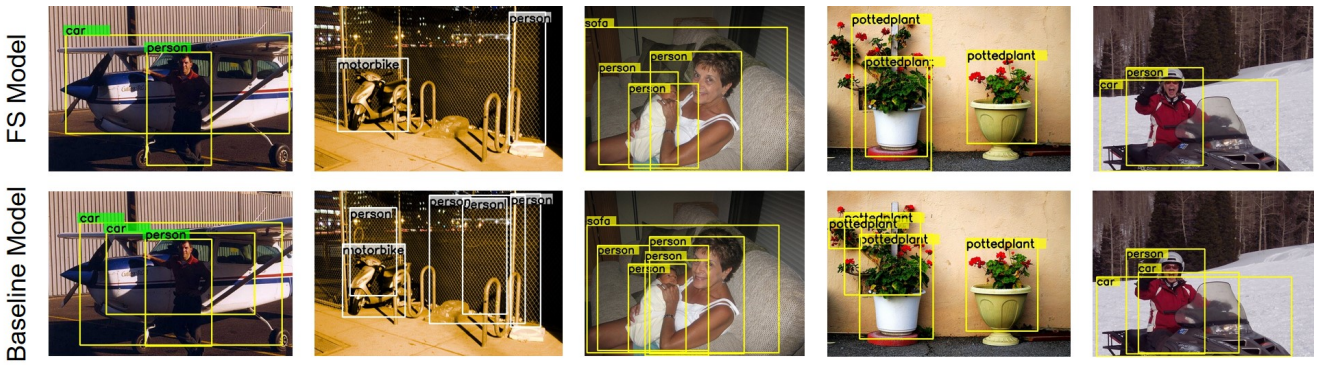

Figure 1: **Examples to explain the label noise overfitting problem.** The first row is pseudo labels from supervised model, and the second row is from semi-supervised model (baseline method). Because of the label noise overfitting problem, pseudo labels from the semi-supervised model contain more similar mistakes.

Because of the label noise overfitting problem, the detector is easy to concentrate on difficult images with more noise and uncertain regions where groundtruth labels are missed, while ignoring easy images and certain regions. Therefore, at *image level*, we propose **a multi-phase learning method guided by uncertainty based image selection** to adopt multiple models handling easy and difficult images separately; at *region level*, we present **a region uncertainty based RoI Re-weighting strategy** to further guide the SSOD training and force the model into more certain regions.

2. Quantitative Results

2.1. Extensive Comparison with Existing Methods

To evaluate our method more comprehensively, we perform more experiments for SSOD on ResNet101 [2] based Faster RCNN [8] and SSD512 [6], two more complicated models compared to ResNet50 based Faster RCNN and SSD300. The more extensive results are listed in Tab. 1 and Tab. 2.

For PASCAL VOC [1], we observe that our method improves the mAP of the baseline method by 3% on ResNet101 based Faster RCNN, from 76.6% to 79.6%, which is consistent with results on ResNet50. This demonstrates that our method can be applied to detectors with different backbones and validates the efficiency of our method. Compared to 78.6% mAP with

Table 1: **Semi-supervised Detection Results on PASCAL VOC 2007 test** *vs.* current SSOD methods and fully-supervised results trained on VOC07 or VOC0712. (L: labeled data, Un: unlabeled data.)

Model	Backbone	Method	L	Un	mAP
Faster RCNN	ResNet50	FS	VOC07	-	74.8
		Baseline	VOC07	VOC12	75.6
		DD [7]	VOC07	VOC12	76.0
		ours	VOC07	VOC12	**78.6**
		FS	VOC0712	-	81.2
	ResNet101	FS	VOC07	-	76.1
		Baseline	VOC07	VOC12	76.6
		DD	VOC07	VOC12	76.9
		ours	VOC07	VOC12	**79.6**
		FS	VOC0712	-	82.2

Model	Backbone	Method	L	Un	mAP
SSD300	VGG16 [9]	FS	VOC07	-	70.2
		Baseline	VOC07	VOC12	71.8
		CSD [3]	VOC07	VOC12	72.3
		ISD [4]	VOC07	VOC12	73.3
		ours	VOC07	VOC12	**74.5**
		FS	VOC0712	-	77.2
SSD512		FS	VOC07	-	73.3
		Baseline	VOC07	VOC12	74.8
		CSD	VOC07	VOC12	75.8
		ISD	VOC07	VOC12	76.4
		ours	VOC07	VOC12	**77.9**
		FS	VOC0712	-	79.6

Table 2: **Semi-supervised detection Results on COCO minival** *vs.* current SSOD and FSOD results.

Backbone	Method	L	Un	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
ResNet50	FS	co-35	-	31.3	52.0	33.0	17.7	34.2	40.0
	DD	co-35	co-80	33.1	53.3	35.4	17.7	36.1	43.4
	ours	co-35	co-80	**34.8**	**55.1**	**37.2**	**19.9**	**37.8**	**45.4**
	ours + DD	co-35	co-80	**35.2**	**55.7**	**37.6**	**20.5**	**38.2**	**45.6**
	FS	co-115	-	37.4	58.1	40.4	21.2	41.0	48.1
	DD	co-115	co-120	37.9	60.1	40.8	20.3	41.6	50.8
	PL [10]	co-115	co-120	38.4	59.7	41.7	22.6	41.8	50.6
	ours	co-115	co-120	**40.1**	**60.4**	**43.7**	**23.6**	**43.7**	**51.4**
	ours + DD	co-115	co-120	**40.3**	**61.0**	**43.9**	**23.9**	**44.0**	**51.5**
ResNet101	FS	co-35	-	32.7	53.4	35.0	17.1	36.2	42.7
	DD	co-35	co-80	34.5	55.7	36.7	19.1	38.8	45.2
	ours	co-35	co-80	**36.4**	**56.5**	**39.2**	**19.5**	**40.1**	**47.7**
	ours + DD	co-35	co-80	**36.6**	**56.8**	**39.4**	**20.1**	**40.3**	**48.0**
	FS	co-115	-	39.4	60.1	43.1	22.4	43.7	51.1
	DD	co-115	co-120	40.1	62.1	43.5	21.7	44.3	53.7
	ours	co-115	co-120	**42.2**	**62.5**	**46.1**	**25.0**	**46.7**	**54.5**
	ours + DD	co-115	co-120	**42.3**	**62.7**	**46.3**	**25.3**	**46.7**	**54.9**

the ResNet50 backbone, ResNet101 achieves a higher mAP and has more potential for practical application. For SSD512, our method also manages to produce better predictions compared to state-of-the-art methods on one stage detectors and our method with SSD300. It is noticeable that the gap between our method (77.9%) and the upper bound by fully-supervised learning on VOC0712 (79.6%) is only 1.7%, which further proves the strong ability of our method.

For COCO [5], we observe that our method persistently improves the mAP on objects with different scales of size, which demonstrates the strong generalization ability of our method. For different backbones or different data splits, our method consistently achieves better performance compared to existing SSOD methods. The most prominent phenomenon is that our method achieves 20.5% on small objects with ResNet50 based Faster RCNN for co-35/80 split, while the corresponding fully-supervised upper bound is just 21.2%. **The gap is only 0.7%.** This demonstrates that our method learns knowledge within unlabeled images quite sufficiently, especially for some certain data.

2.2. Image Difficulty Distribution

With recall/precision metric we propose in the original paper, the fraction of easy images is calculated and listed in Tab 3. For ResNet50 based Faster RCNN on PASCAL VOC, the fraction of easy images is 54%, which is consistent with the result in our original paper where 50% or 60% easy images produce a higher mAP. If we combine our method with DD or use ResNet101 as the backbone, the quality of pseudo labels is higher, but we find that the easy data ratio is still close to 50%. For more complicated datasets like COCO, the proportion of easy images is still approximately 50%. This confirms

Table 3: **Fraction of easy data proportion** on PASCAL VOC and MS COCO

Dataset	Model	Backbone	DD	Easy Data Proportion
PASCAL VOC	Faster RCNN	ResNet50		54%
			✓	61%
		ResNet101		59%
			✓	64%
MS COCO	Faster RCNN	ResNet50		46%
			✓	53%
		ResNet101		50%
			✓	56%

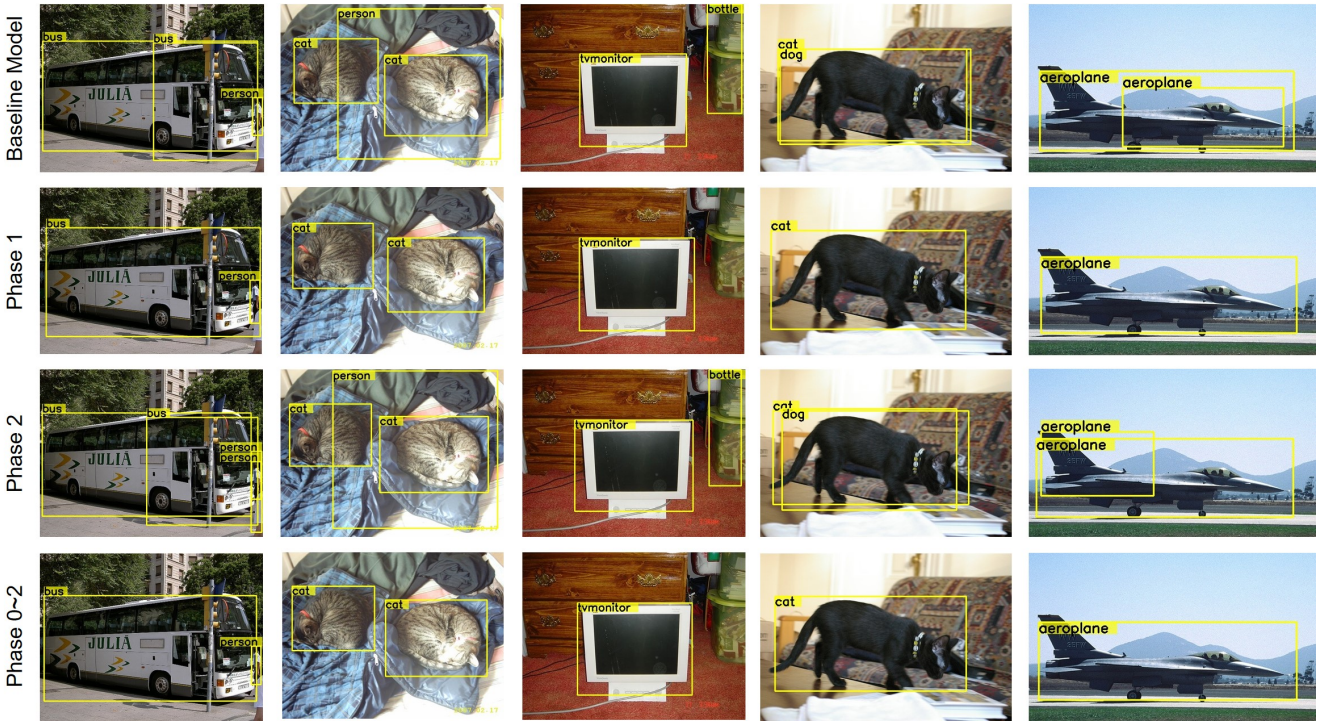

Figure 2: **Comparative examples on easy images.** The first row is from the baseline method, the second row is from the phase 1 model, the third row is from the phase 2 model, and the fourth row is from the phase $0 \sim 2$ model. The phase 1 model performs the best on these easy images, while the baseline model and the phase 2 model concentrates on difficult knowledge and makes mistakes on easy images.

that without any prior knowledge about the unlabeled dataset, 50% is a good estimation of easy data.

3. Qualitative Results

3.1. Overall Visualized Results

Fig. 2 and Fig. 3 shows the visualized results of baseline methods and our methods on several easy or difficult images.

For easy images, we observe that the common points among them are that the objects within them are usually large and clear. The features for those objects are also obvious and they are not easy to be confused with other categories. We find that the detector from the first phase performs the best because they are trained with only easy unlabeled images and excel in easy knowledge. The baseline method and the second phase model both detect some extra objects that do not exist. This is because **these two models are trained with all unlabeled images and focus on difficult knowledge**. They thus keep a high vigilance and perform prediction even the corresponding region is only a little similar to those objects, such as the right

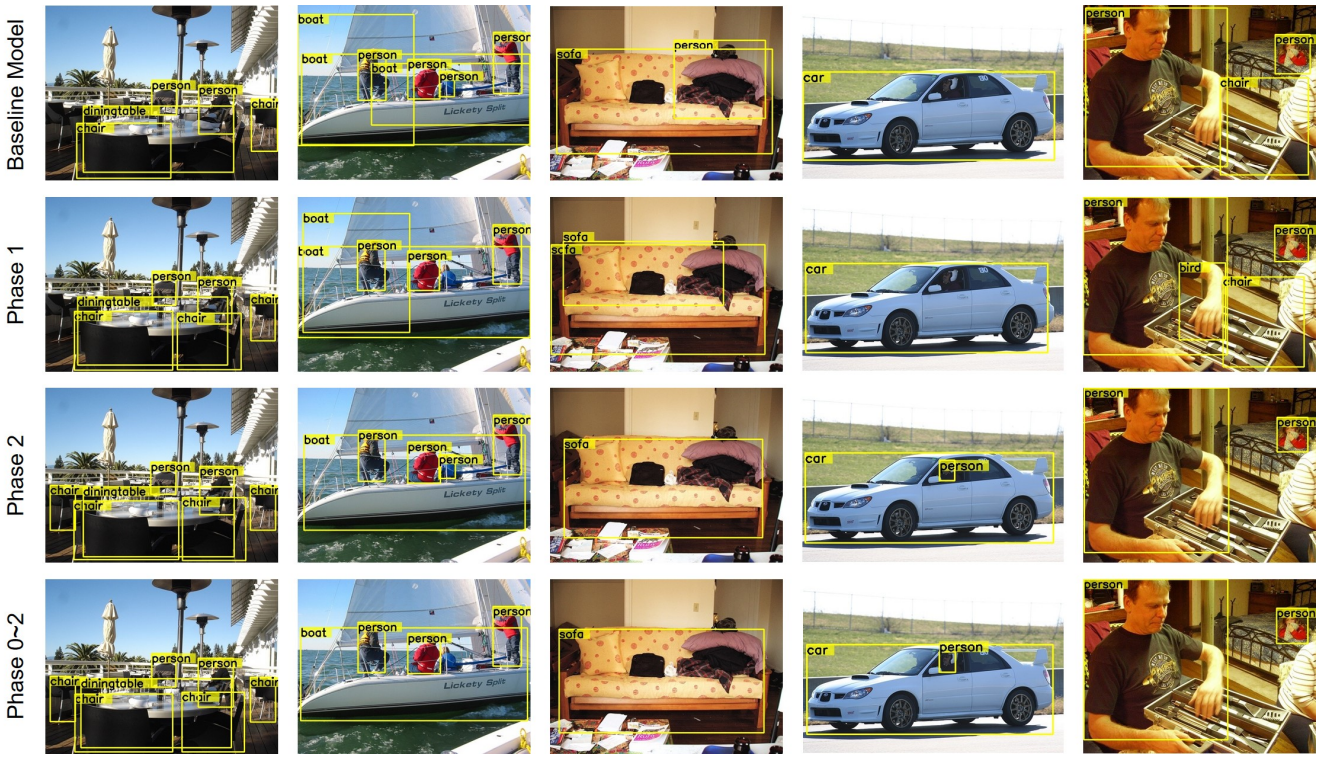

Figure 3: **Comparative examples on difficult images.** The first row is from the baseline method, the second row is from the phase 1 model, the third row is from the phase 2 model, and the fourth row is from the phase $0 \sim 2$ model. The phase 2 model performs the best on these difficult images, while the phase 1 model concentrates on easy knowledge and is not able to detect some confusing objects.

'bottle' in the third case. They are also easy to detect parts of the objects, such as the first and the fifth case. These images demonstrate that **the model from the first phase experts in easy knowledge.** After ensembling, the confidence scores of those redundant predictions are lowered down and they are not distinct in the final prediction.

For difficult images, we find that the number of objects within them is a little large, or they contain some small and unclear objects. Some objects are easy to be confused with other categories or the background category. For example, the clothes in the third case look like a person, and the right small person in the fifth case is hard to distinguish from the environment. For these images, the second phase model is the best. The first phase model either makes some confusion or is not able to detect some objects. These images illustrate that **the model from the second phase learns difficult knowledge well.**

After ensembling, the model takes the advantage of all models thus is suitable for both easy and difficult images. Therefore, our method achieves a better result compared to the baseline method.

3.2. Visualized Results for RoI Re-weighting

We list more examples in Fig. 4. We observe that our method also manages to detect uncertain regions where groundtruth labels are miss-annotated, such as the middle person for the first case, the bottles in the second case, the right cow in the third case, the right person in the fourth case, and the left dog in the fifth case. **Weights for these regions are reduced and incorrect gradient information deriving from uncertain regions is less likely to negatively affect the detector.**

From Fig. 5, we observe that with RoI Re-weighting, our model is able to detect more confusing and difficult objects compared to the model without RoI Re-weighting. This is just because that RoI Re-weighting helps the model avoid uncertain regions. Our model is less guided by uncertain and incorrect knowledge and is thus able to detect more difficult objects.

References

[1] 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. 1

Figure 4: **Illustrative examples for RoI Re-weighting.** The first row is pseudo labels with missing annotation problem and the second row is heatmap for region uncertainty. Our method manages to detect uncertain regions (blue ones).

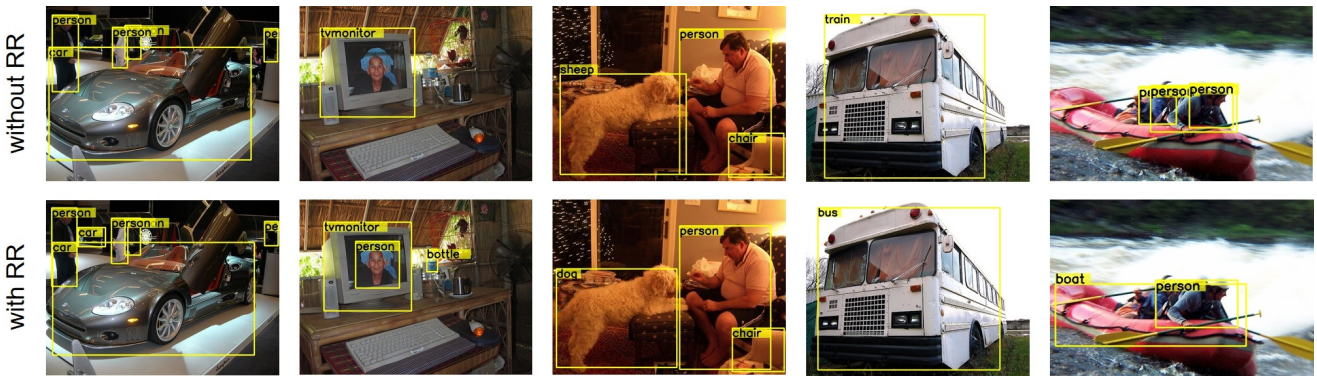

Figure 5: **Comparative examples for RoI Re-weighting.** The first row is from the two-phase model without RoI Re-weighting, and the second row is from the two-phase model with RoI Re-weighting. Because of RoI Re-weighting, the detector is able to handle more confusing objects.

[2] 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. 1

[3] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. In *Advances in neural information processing systems*, pages 10759–10768, 2019. 2

[4] Jisoo Jeong, Vikas Verma, Minsung Hyun, Juho Kannala, and Nojun Kwak. Interpolation-based semi-supervised learning for object detection. *arXiv preprint arXiv:2006.02158*, 2020. 2

[5] 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. 2

[6] 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. 1

[7] 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

[8] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1

[9] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[10] Peng Tang, Chetan Ramaiah, Ran Xu, and Caiming Xiong. Proposal learning for semi-supervised object detection. *arXiv preprint arXiv:2001.05086*, 2020. 2