Gradient-based Sampling for Class Imbalanced Semi-supervised Object Detection (Supplementary Document)

Anonymous ICCV submission

Paper ID 5153

1. Choice of Hyperparameters

We analyze the choice of two hyper-parameters in our approach, γ, and β.

Choice of β. We introduce β to prevent the model from overfitting to the small number of labeled data of minority classes. In the training process, the detector is prone to overfit fixed annotations of minority classes. By applying large weights, the predictions for minority classes on the labeled set could be over-confident. Consequently, the positive gradients of these classes become smaller and enforce the class-wise weights to become larger, leading to more serious overfitting. However, this issue is trivial in the unlabeled set. When the weights for a minority class become larger, more pseudo labels are sampled for training, which increases the gradient for this minority class to prevent the model from overfitting. Besides, since the sampling rate of labeled images is always larger than unlabeled images in current SSOD methods, the model tends to overfit with labeled images. Small class weights for labeled data can effectively alleviate overfitting. Thus, we smooth the class weights for labeled images when keeping the weights for unlabeled data the same to balance the positive gradient and negative gradient of each class. We compare the performance of the detector under different β in COCO split1 in Tab.1. When β is set to 1, the performance of minority classes is better than that under $\beta = 0.5$ by 1.5%, but the performance of majority classes drops by 1.4%. It demonstrates that a larger β is beneficial to minority classes but causes underfitting of majority classes.

Choice of γ. Tab.2 illustrates the AP of our detector under different values of γ. In general, our method is not sensitive to the choice of γ. Setting γ to 0.5 leads to a slight performance improvement in minority classes but causes a small performance drop in majority classes. In fact, γ is a parameter depending on the class frequencies of the dataset. As the mean frequency of all classes in MS-COCO $\rightarrow$ Object365 and LVIS sub-tasks is much smaller, we set a smaller γ for these tasks.

Table 1: Performance comparisons of setting different β on the split1 of MS-COCO sub-task. AP is chosen as the metric.

β	all	maj	min
0.3	**26.5**	**30.8**	13.4
0.5	**26.5**	30.7	13.9
1	25.8	29.3	**15.4**

Table 2: Performance comparisons of setting different γ on the split1 of MS-COCO sub-task. AP is chosen as the metric.

γ	all	maj	min
0.1	26.4	**30.7**	13.5
0.5	**26.5**	**30.7**	**13.9**
1	26.3	30.4	14.0

2. Motivation of Jacobi Iterate Module

By optimizing the equations of

$$w_i \times \tilde{\boldsymbol{G}}_{ii} = -\sum_{k \neq i} w_k \times \tilde{\boldsymbol{G}}_{ik} \tag{1}$$

and

$$\sum_{i=1}^{n+1} w_i = n + 1 \tag{2}$$

we obtain a class weights $w_i (i = 1, 2, ..., n)$ utilized for the reweighting and thresholding in our method. However, traditional direct solutions to optimize the group of linear equations always obtain a negative weight $w_i < 0$ in complex tasks such as MS-COCO $\rightarrow$ Object365 and LVIS subtasks. To solve this, we introduce the learnable parameters $a_i (i = 1, 2, ..., n + 1)$ with softmax function to obtain w_i subject to $w_i > 0$. To update the value of a_i, we introduce a loss to mimic the Jacobi iterate method. Experimentally,

Table 3: Performance comparisons of using different sampling strategies on the MS-COCO sub-task. AP is chosen as the metric.

labeled	unlabeled	all	maj	min
Random	Random	25.5	**30.9**	9.4
RFS	Random	25.8	**30.9**	10.5
RFS	RFS	26.0	30.8	11.7
RFS	CrS	**26.5**	30.7	**13.9**

the AP of the detector trained with our Jacobi iterate module (26.5%) is the same as that of the detector trained with direct solution (26.5%) in the MS-COCO sub-task.

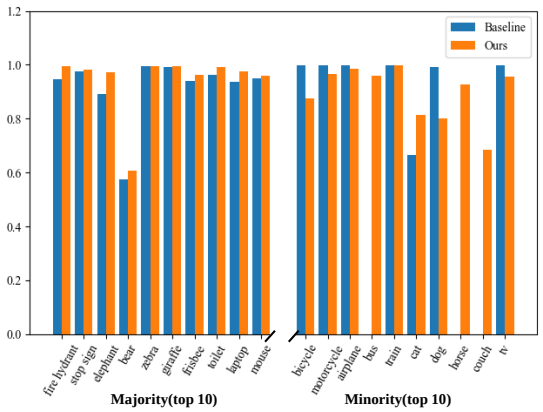
3. Extended Ablation of Sampling Strategies

We conduct experiments on split1 of the MS-COCO subtask to further demonstrate the effectiveness of our proposed CrS module. Different sampling strategies are applied to labeled and unlabeled images. According to Tab. 3, our method performs best among all variants. Concretely, we choose the random sampling strategy applied to both labeled and unlabeled data as the baseline, which achieves 30.9% and 9.7% on the majority classes and the minority classes, respectively. Changing the random sampling strategy to RFS increases the performance by 0.4% on all classes, and we choose this RFS variant as our strong baseline. By replacing the RFS on unlabeled data with our proposed CrS module, the performance of minority classes increases by 2.2%, and the performance of all classes also further increases by 0.5%. This demonstrates the superiority of our CrS module over RFS on unlabeled data.

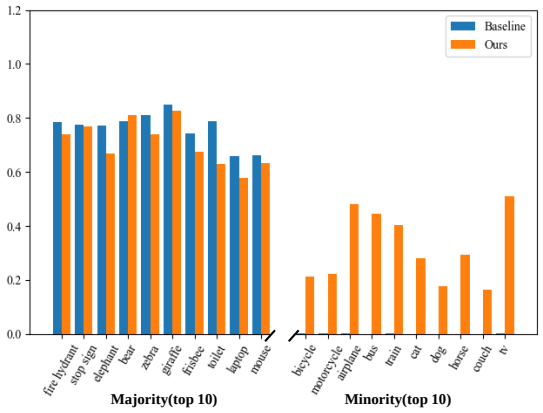
4. Analysis of Confirmation Bias

Confirmation bias[1] refers to the phenomenon that the model training is overfitted to the incorrect pseudo labels predicted by the model itself. Here, we present an analysis of confirmation bias in our task. We analyze the confirmation bias in our tasks from two perspectives: (1) the model biased toward majority classes tends to predict pseudo labels biased toward majority classes; (2) the incorrect pseudo labels for minority classes dominate the learning of the model since the ground truth labels in labeled data are scarce for these classes.

To illustrate the first aspect of this phenomenon, we select Top-10 AP classes for both majority classes and minority classes and calculate the precision and recall metrics of pseudo labels of Top-10 classes in both majority classes and minority classes. The results are presented in Fig. 1. Obviously, there is a performance bias towards the majority classes of the baseline (Soft Teacher), which can be seen from the recall metric in Fig. 1 (b). This manifests that the detector is leaning towards generating more pseudo labels



(a) Precision of pseudo labels

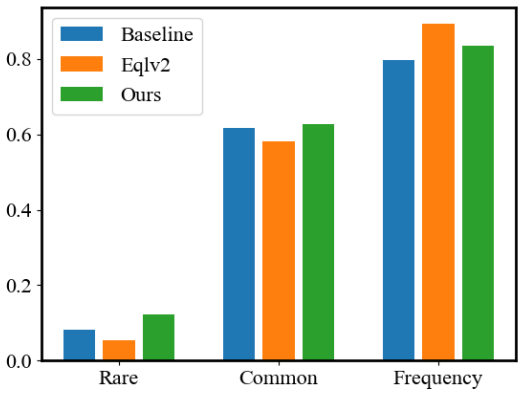


(b) Recall of pseudo labels

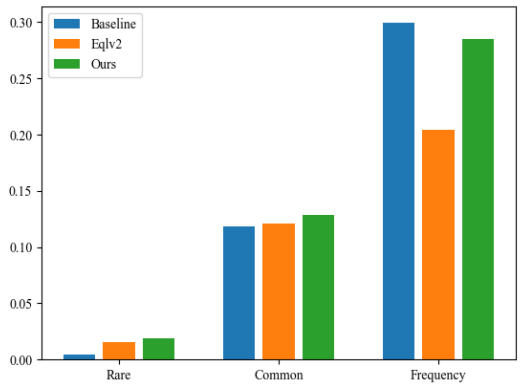
Figure 1: Comparisons between our method and the baseline (Soft Teacher) regarding the precision and recall metrics of pseudo labels of each class in MS-COCO sub task. We select Top-10 classes for majority classes and minority classes based on the AP performance, respectively.

for the majority classes while ignoring the minority classes. However, our proposed framework achieves a better trade-off between the precision and recall metrics of both majority classes and minority classes. Specifically, our method has a higher recall of pseudo labels for minority classes and preserves the high precision and recall of pseudo labels of majority classes. This demonstrates that our method can effectively alleviate the confirmation bias towards majority classes.

To illustrate the second aspect of the confirmation bias, we analyze the precision and recall metrics of pseudo labels in LVIS sub-task for Soft Teacher(baseline), the Long-tailed method EqlV2[7], and our method. In this task, the implicit instances of minority classes in unlabeled data are always



(a) Precision of pseudo labels



(b) Recall of pseudo labels

Figure 2: Comparisons between the baseline (Soft Teacher), EqlV2 and our method regarding the precision and recall metrics of pseudo labels of each class in LVIS sub task.

scarce. The pseudo labels for minority classes are prone to be incorrect and thus introduce noises for the detector in the training process. The reasons are two-fold: (1) Since the ground truth annotations for these classes are few, it is hard for the model to learn sufficient knowledge for these classes to generate high-quality pseudo labels. (2) Different from the classification task, there are lots of instances of other classes and uncorrelated instances(regarded as background, for example) in an image. As ground truth labels and the implicit instances in unlabeled images are scarce, the learning of detectors is heavily interfered with by incorrect pseudo labels from instances from other classes and uncorrelated instances. As shown in Fig. 2, the precision of pseudo labels of rare classes is much lower than that of common and frequency classes. It demonstrates that most of pseudo labels for minority classes are incorrect and thus lead to confirmation bias. Although EqlV2 generates more pseudo labels for minority classes than the Baseline, the precision of these pseudo labels is lower. In contrast, our method generates more pseudo labels for minority classes with high precision, showing superiority to combat the confirmation bias.

5. Model Analysis

We present more model analysis by answering the following two questions: (1) Do we have more balanced gradients in training? (2) Do we have better representations?

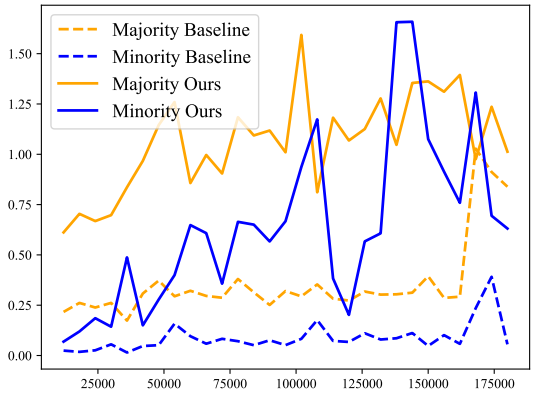
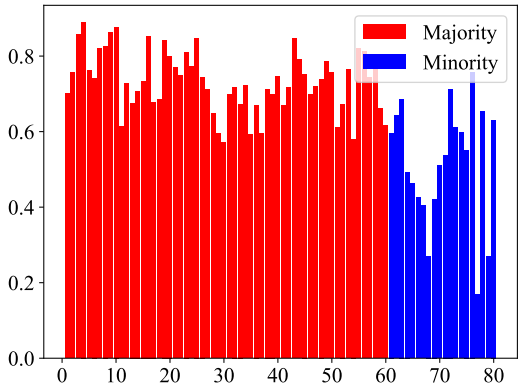


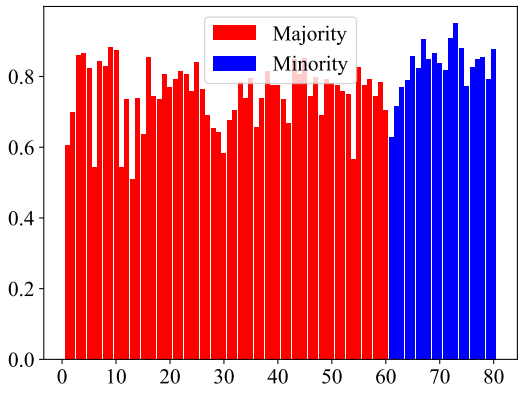
Figure 3: Comparisons between our method and the baseline (Soft Teacher) regarding the averaged gradient norm of the majority classes and that of the minority classes during the training process. The dashed dot lines denote the baseline model and the solid lines denote our method. The YELLOW lines represent the mean gradient norm of the majority classes and the BLUE lines denote the mean gradient norm of the minority classes.

Do we have more balanced gradients in training? We present the comparisons between our method and the baseline (Soft Teacher) regarding the averaged gradient norm of the majority classes and that of the minority classes during the training process in Fig. 3. According to Fig. 3, it is clear that (1) our method exhibits a higher gradient norm than that of the baseline in both majority and minority classes; (2) as the training proceeds, our method shows that the gradient norm gap between the majority classes and the minority classes becomes smaller while the baseline model does not show such a property. These two observations manifest that our method has achieved a more balanced gradient in training. Besides, we further present the positive and negative gradient norms for a representative minority class(laptop) in Fig 5, which shows that the positive and negative gradi-

ent norms can be **equalized** with GbR.



(a) Baseline



(b) Ours

Figure 4: Comparisons between our method and the baseline (Soft Teacher) regarding the mean prediction score of each class.

Do we have better representations? We adopt the mean prediction score as the metric to measure whether the representation of a class is better or not. That is, the higher the averaged prediction scores for a class, the better the representation of that class is. Thus, we visualize the mean prediction score of each class when the model is converged. According to Fig. 4, for the majority classes, the mean prediction scores of both our method and the baseline are almost the same, suggesting that they both have good representations; for the minority classes, it is obvious that the prediction scores of our method are much higher than that of the baseline, demonstrating that our method has better representations for the minority classes which result in a class balanced object detector. Besides, as shown in Fig 6, we also visualize the class logits with or without our method. The class logits with GbR are much more balanced than logits without GBR. It shows our method can level off the class logits during training to learn better representations.

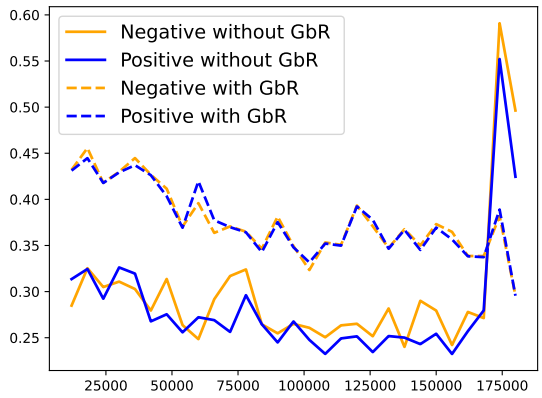


Figure 5: The visualization of positive and negative gradients with and without GbR.

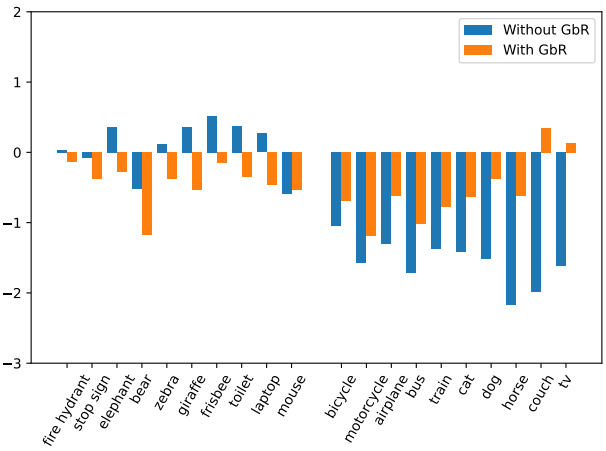


Figure 6: The comparison of class logits with and without GbR.

6. More Implements Details

To stabilize the gradient statistics in the gradients matrix $\tilde{\boldsymbol{G}}$, we propose to employ a warm-up process before solving the analytical solutions for the class-wise weights $w_i (i = 1, ..., n + 1)$. Specifically, we adopt 4000 iterations for the MS-COCO sub-task and 16000 iterations for the MS-COCO $\rightarrow$ Object365 and LVIS sub-task. During the warm-up process, the Gradient-based Reweighting(GbR) module and the Gradient-based Reweighting(GbT) module

Table 4: Performance of our method with Unbiased Teacher[3] on the split1 of MS-COCO sub-task

Methods	all	maj	min
Unbiased Teacher	22.3	28.6	5.3
+ Ours	24.7	28.5	13.2

Table 5: Comparisons with long-tailed learning methods on the split1 of MS-COCO sub-task.

Methods	all	maj	min
Logits Adjustments	24.6	30.6	6.5
Classifier Retraining	24.5	29.9	8.4
Ours	**26.5**	**30.7**	**13.9**

are not turned on in training. That is $w_i = 1$ and $\theta_i = \theta$ for $i = 1, 2, ..., n+1$. We will also release our dataset and code in the future to show more details.

7. Experiments on more SSOD detectors

We also conduct the experiments with Unbiased Teacher[3] in Tab 4. The results demonstrate the generalization of our method on different SSOD detectors.

8. Comparisions with more long-tailed learning methods

We compare our method with Logits Adjustments[4] and Classifier Retraining[2] in Tab 5. It shows our method beat existing long-tailed learning methods by a clear margin, demonstrating the effectiveness of our method.

9. Extended Descriptions of Gradient-based Reweighting

We present a simple example to elaborate on the calculation of the gradient matrix $\tilde{\boldsymbol{G}}$ shown in Fig. 7. The class-wise weights are the analytical solutions of the linear equations regarding $\tilde{\boldsymbol{G}}$. The calculated weights are then applied to the classification losses $f^l(\cdot)$ and $f^u(\cdot)$ to achieve class-balanced training at the gradient level and to the Gradient-based Thresholding module to acquire more pseudo labels for the minority classes at the image-level.

10. Qualitative Comparisons against Other Methods

We present several examples of qualitative comparisons between our method and current methods, including Soft Teacher[9], C2AM loss[8], DASO[5] and DeFRCN[6] shown in Fig. 8 for the MS-COCO sub-task, Fig. 9 for the MS-COCO $\rightarrow$ Object365 sub-task and Fig. 10 for the LVIS sub-task. Obviously, our method generates more detection results for minority classes while keeping the same amount of detection results for the majority classes, leading to a better overall performance improvement on both sub-tasks.

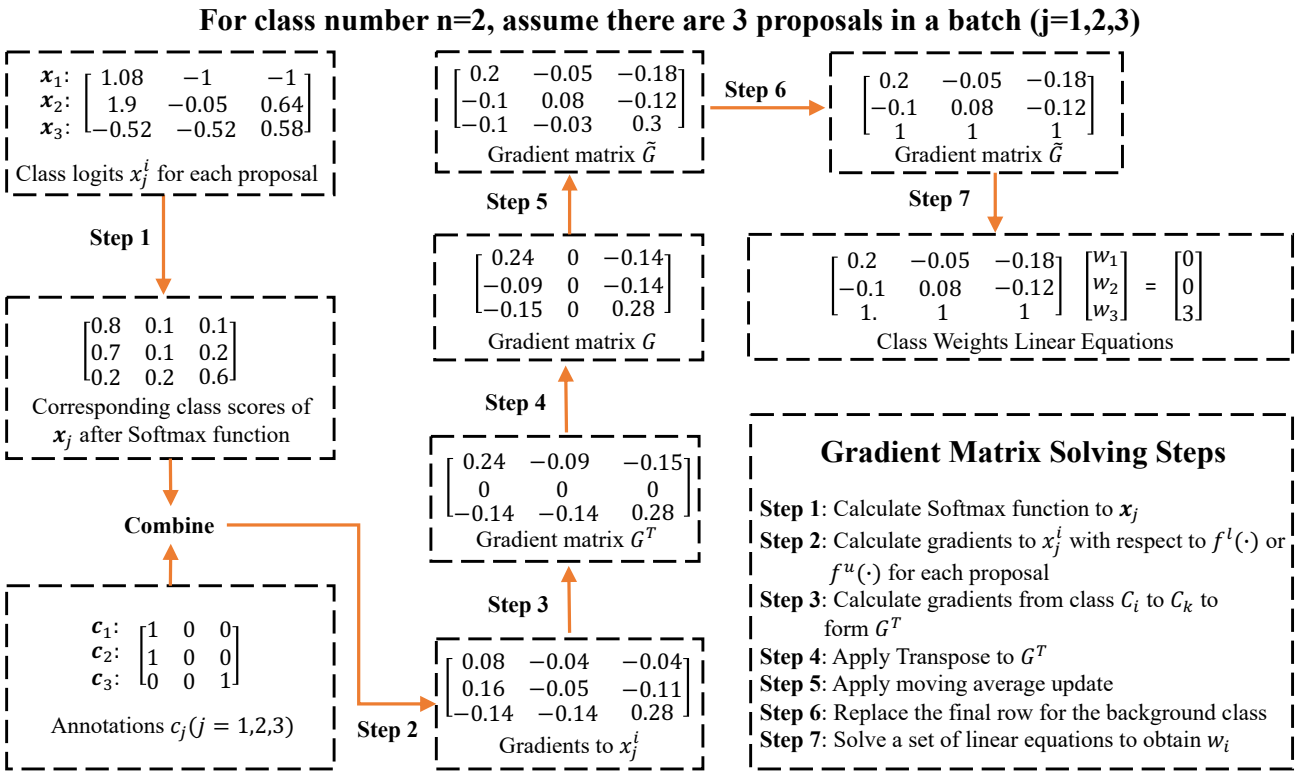


Figure 7: A simple example to illustrate the process of Gradient-based Reweighting.

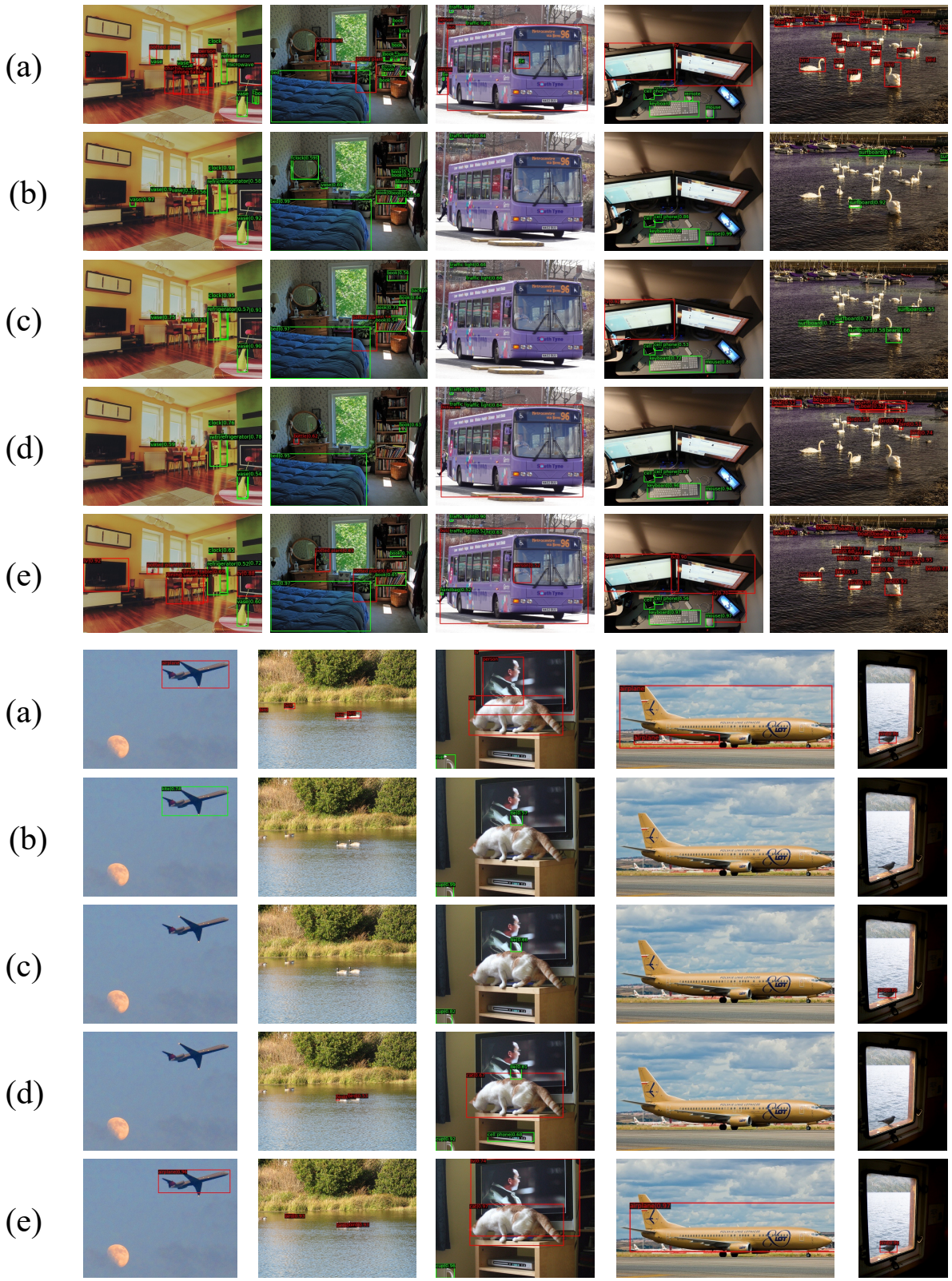


Figure 8: Qualitative comparisons against other methods in MS-COCO sub-task:(a) Ground Truth labels; (b) Soft Teacher[9]; (c) DASO[5]; (d) DeFRCN[6]; (e) Ours. The bounding boxes with RED borderline are predicted as minority classes while the bounding boxes with GREEN borderline are predicted as majority classes.

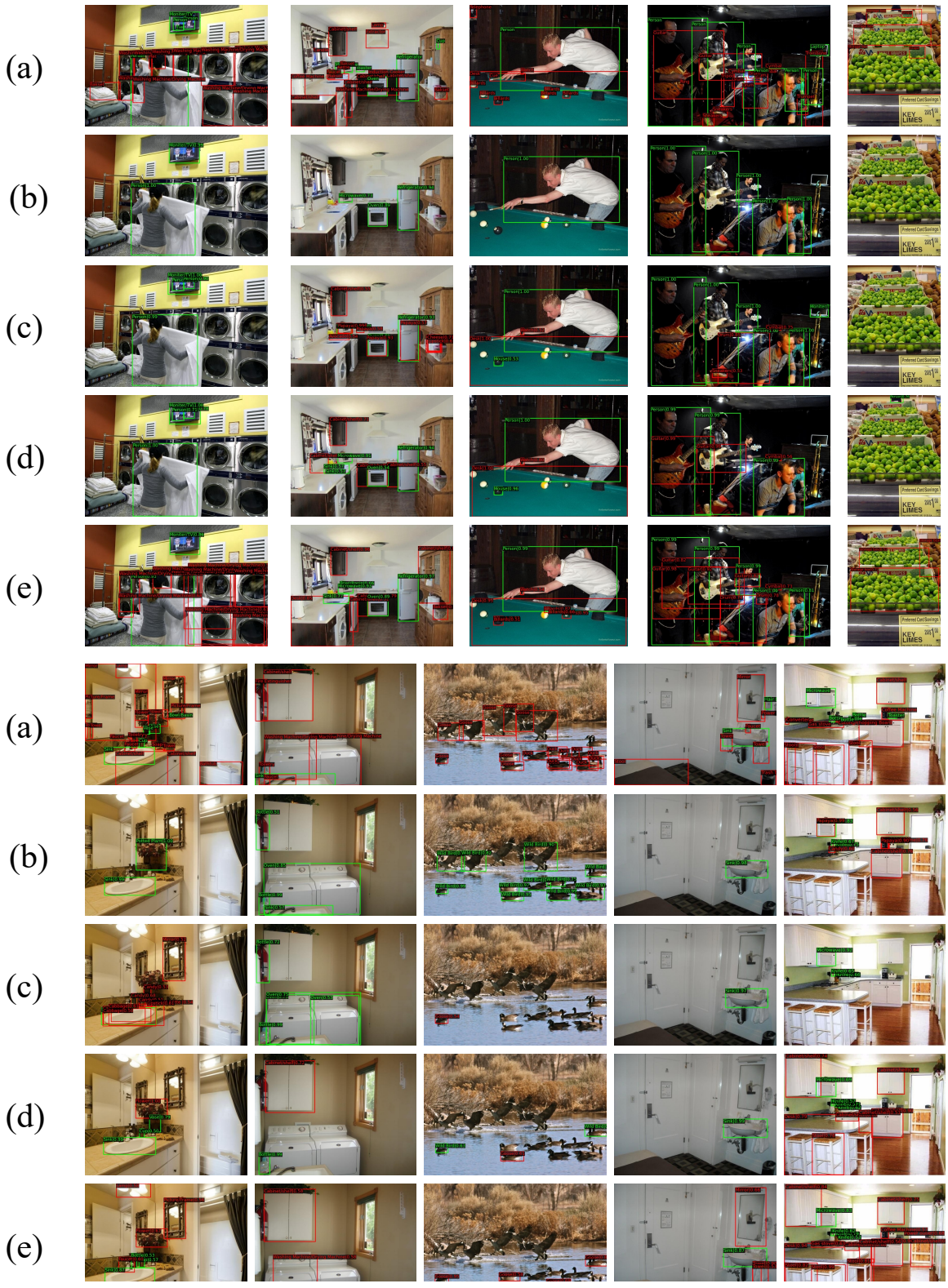


Figure 9: Qualitative comparisons against other methods in MS-COCO → Object365 sub-task:(a) Ground Truth labels; (b) Soft Teacher[9]; (c) Eqlv2[7]; (d) DeFRCN[6]; (e) Ours. The bounding boxes with RED borderline are predicted as minority classes while the bounding boxes with GREEN borderline are predicted as majority classes.

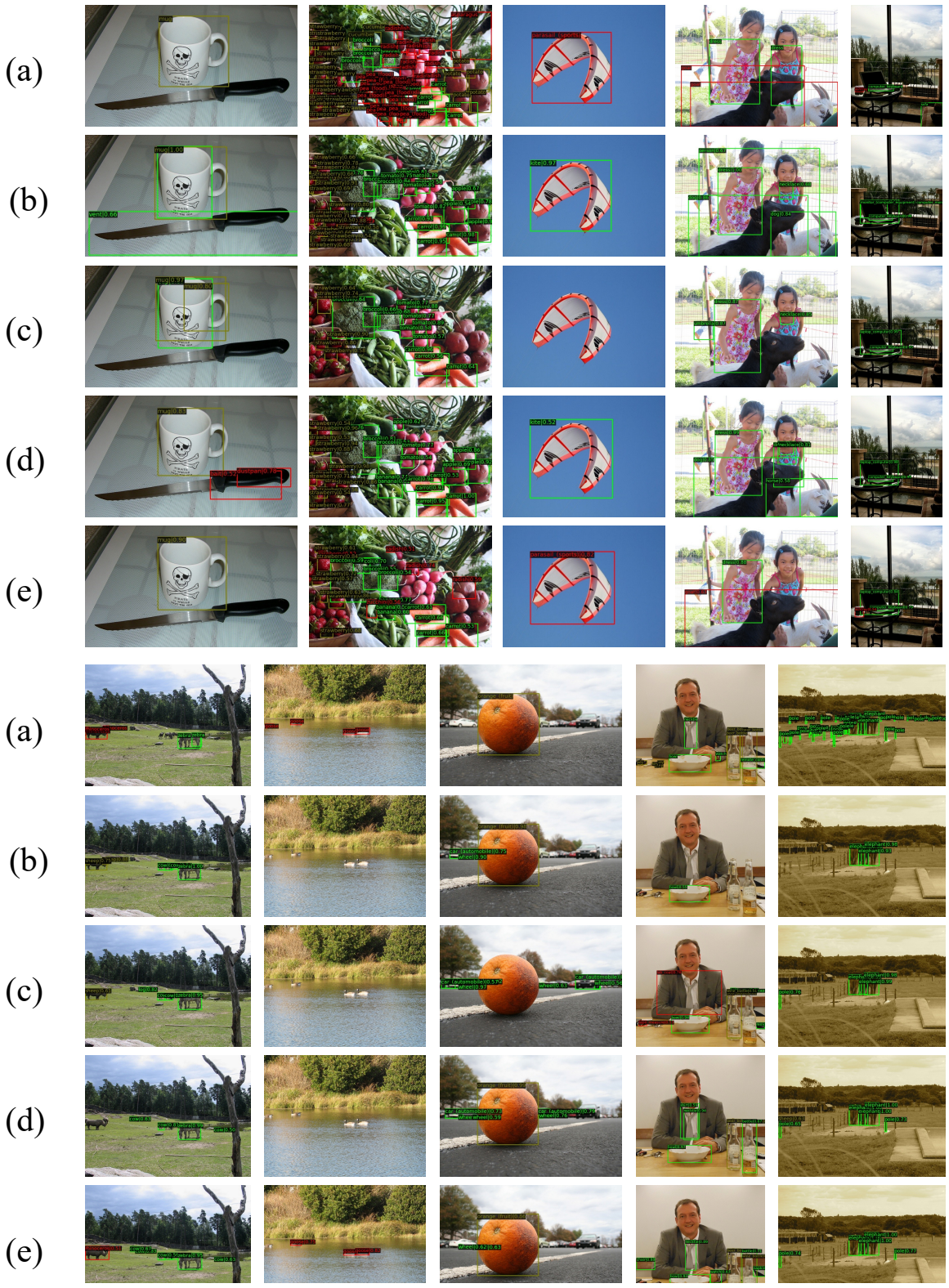


Figure 10: Qualitative comparisons against other methods in LVIS sub-task:(a) Ground Truth labels; (b) Soft Teacher[9]; (c)Eqlv2[7]; (d) DASO[5]; (e) Ours. The bounding boxes in RED, YELLOW and GREEN are predicted as rare classes, common classes and frequent classes, respectively. Note that the instances in LVIS dataset are not fully annotated.

References

[1] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020. 2

[2] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019. 5

[3] 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. 5

[4] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020. 5

[5] Youngtaek Oh, Dong-Jin Kim, and In So Kweon. Daso: Distribution-aware semantics-oriented pseudo-label for imbalanced semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9786–9796, 2022. 5, 7, 9

[6] Limeng Qiao, Yuxuan Zhao, Zhiyuan Li, Xi Qiu, Jianan Wu, and Chi Zhang. Defrcn: Decoupled faster r-cnn for few-shot object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8681–8690, 2021. 5, 7, 8

[7] Jingru Tan, Xin Lu, Gang Zhang, Changqing Yin, and Quanquan Li. Equalization loss v2: A new gradient balance approach for long-tailed object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1685–1694, 2021. 2, 8, 9

[8] Tong Wang, Yousong Zhu, Yingying Chen, Chaoyang Zhao, Bin Yu, Jinqiao Wang, and Ming Tang. C2am loss: Chasing a better decision boundary for long-tail object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6980–6989, 2022. 5

[9] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3060–3069, 2021. 5, 7, 8, 9