Model	AP	AP_r	AP_c	AP_f
Challenge Baseline	30.1	19.3	31.8	32.3
+SE154 [3]	30.8	19.7	32.2	33.4
+OpenImage Data	31.4	21.5	33.1	33.3
+Multi-scale box testing	32.3	20.5	34.7	34.2
+RS Ensemble+Expert Model	35.1	24.8	37.5	36.3
+Multi-scale mask testing	**36.4**	**25.5**	**38.6**	**38.1**

Table 1: Experiment results of different tricks on LVIS v0.5 `val` set. RS Ensemble stands for Re-scoring Ensemble.

Appendix A. Details of LVIS Challenge 2019

With equalization loss, we ranked 1st entry on LVIS Challenge 2019. In this section, we will introduce details of the solution we used in the challenge.

External Data Exploiting Since LVIS is not exhaustively annotated with all categories and the annotations for long-tailed categories are quite scarce, we utilize additional public datasets to enrich our training set. First, we train a Mask R-CNN on COCO train2017 with 115k images and then fine-tune our model with equalization loss on LVIS. During fine-tuning, we leverage COCO annotations of bounding boxes as ignored regions to exclude background proposals during sampling. Moreover, we borrow $\sim$20k images from Open Images V5 which contains shared 110 categories with LVIS and use the bounding boxes annotations to train the model.

Model Enhancements We achieve our challenge baseline by training ResNeXt-101-64x4d [5] enhanced by deformable convolution [1] and synchronized batch normalization [4], along with equalization loss, repeat factor sampling , multi-scale training and COCO exploiting, which lead to 30.1% AP on LVIS v0.5 `val` set. We apply multi-scale testing on both bounding box and segmentation results and the testing scale ranges from 600 to 1400 with step size of 200. We train two expert models on train set of COCO 2017 and Open Images V5 respectively and then evaluate them on LVIS `val` set to collect the detection results of shared categories. Though our method improves the performance of long-tailed categories a lot, the prediction scores for these categories tend to be smaller than frequent ones due to the lack of positive training samples, which leads to degeneration of AP_r in ensemble. To keep more results for rare and common categories, we employ a re-score ensemble approach via improving the scores of these categories.

Our road map is shown in 1. With those enhancements, we achieve 36.4 and 28.9 Mask AP on `val` and `test` set respectively which is demonstrated in Table 2.

	AP	AP_{50}	AP_{75}	AP_r	AP_c	AP_f
Ours	**28.85**	**42.69**	**31.10**	**17.71**	**30.84**	36.70
2nd place	26.67	38.68	28.78	10.59	28.70	**39.21**
3rd place	24.04	36.51	25.68	15.32	24.95	31.12
4th place	22.75	34.29	24.17	11.67	23.51	32.33
Baseline [2]	20.49	32.33	21.57	9.80	21.05	30.00

Table 2: Results reported on LVIS v0.5 20k `test` set. The equalization loss plays a important role to allow us to achieve the highest AP both on rare and common categories. This results can be accessed by `https://evalai.cloudcv.org/web/challenges/challenge-page/442/leaderboard/1226`

References

[1] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017.

[2] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019.

[3] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[4] Chao Peng, Tete Xiao, Zeming Li, Yuning Jiang, Xiangyu Zhang, Kai Jia, Gang Yu, and Jian Sun. Megdet: A large mini-batch object detector. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6181–6189, 2018.

[5] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.