Improving the Intra-class Long-tail in 3D Detection via Rare Example Mining

Chiyu Max Jiang, Mahyar Najibi, Charles R. Qi,
Yin Zhou, and Dragomir Anguelov

Waymo LLC., Mountain View CA 94043, USA
{maxjiang,najibi,rqi,yinzhou,dragomir}@waymo.com

1 Experiment Details

We provide additional details for our experiments below.

1.1 Training the Normalizing Flow Model

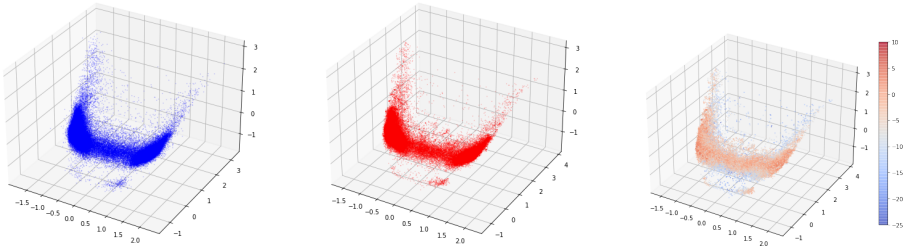
We train a continuous normalizing flow model (FFJORD [1]) for estimating the rareness of detection objects. As detailed in the main text, we train the flow model on the normalized feature vectors after PCA transformation obtained via ROI max-pooling the final feature map of the MVF detector using predicted or ground-truth bounding boxes.

The feature vectors are reduced to the dimension of 10 after the PCA transformation. The flow model consists of a stack of 4 consecutive FFJORD bijectors. Each FFJORD bijector consists of 4 hidden layers, each hidden layer consists of 64 units, and uses `tanh()` activation. The model is trained using an Adam Optimizer, with an initial learning rate of `1e-4`, learning rate decay every 2400 steps, with a decay rate of 0.98. We train the flow model for a total of 100 epochs. For the base distribution of the flow model, we use a spherical Gaussian distribution with unit variance. See Fig. 1 for a visualization of the inferred feature densities using a flow model trained using the procedure above.

1.2 3D Auto-labeler Implementation

We adopt the auto-labeling pipeline as outlined in [2]. We use a strong teacher model to serve as the auto-labeler. For the active learning experiments, we train a 5-frame MVF model [5] on the same 10% fully-labeled segments as the single-frame MVF student model. We then use this 5-frame MVF model to extract boxes, create object tracks using [4], and further refine the detections using the refiner [2].

2 Additional Visualizations of Rare Examples



(a) PCA projection of the training embeddings.

(b) PCA projection of generated embeddings.

(c) Estimated log probability of embeddings.

Fig. 1: Visualization of (a) the embeddings used for training the normalizing flow model, (b) the generated embeddings from the flow model by sampling the learned distribution, that matches the training distribution very well, signifying a good learned estimation of feature densities.(c) a visualization of the estimated log probability of the embeddings from a learned normalizing flow model. The model is able to assign higher log probability to denser features (common examples) and lower log probability to sparser features (rare examples).

Fig. 2: Additional visualizations of rare examples as inferred by the trained flow model. Warmer colors indicate more rare instances in the scene. The most rare instances in the scene include large vehicles (construction vehicles, trucks), vehicles of irregular geometries such as a flatbed trailer, as well as potentially mislabeled objects such as cones.

3 Additional Experimental Evaluation

3.1 Pedestrian Rare Example Mining

We perform an additional experimental evaluation for rare example mining performance on pedestrians to offer more insights into using rare example mining for other object categories. Similar to the main experiments in the paper, we perform this experiment on the Waymo Open Dataset [3]. We use the same MVF [5] backbone as the main pedestrian detector and utilize [2] for auto-labeling the unlabeled instances. We mainly compare three sets of experiments. We compare two active learning experiments that train on 10% of the fully-labeled segments, plus an additional 10% of mined pedestrian tracks from the rest of the training segments, where the tracks are either mined using our MD-REM++ method or by randomly selecting from the remaining tracks. For our MD-REM++ method, we use a hard example filtering function where the number of points threshold $\tilde{p} = 10$ and $d = 50$(m). We report on Average Precision (AP) metrics at an IoU threshold of 0.5. Unlike vehicle experiments where we have an intuitive estimate of rareness based on vehicle size, for pedestrian experiment we instead evaluate the model performance on the intra-class long-tail by evaluating the AP performance on the top-5% rarest ground truth objects (based on inferred log probability). We present our results in Table 1.

Experiment	Human Labels (%)	All	Rare (Top 5%)
Fully Supervised	(100%;0%)	0.757	0.111
Random	(10%;10%)	0.705	0.038
Ours (MD-REM++)	(10%;10%)	**0.729**	**0.060**

Table 1: Active learning experiments for pedestrians on Waymo Open Dataset.

We conclude two main findings from the pedestrian experiment that is consistent with the vehicle experiments:

1. Our rare example mining method is able to significantly outperform baselines.
2. Our flow-based method is able to find challenging instances for the model, as the model performs much worse on rare subsets, compared to the general distribution.

3.2 Additional Evaluation for Vehicle Rare Example Mining

We show additional evaluation for the experiments presented in Table 3 (main paper). We present the results below.

Experiment	Human Labels (%)	All	Regular	Large	Rare (Top 5%)
Fully Supervised	(100%;0%)	0.730	0.732	0.458	0.178
Ours (MD-REM++)	(10%,3%)	0.732	0.735	0.423	0.126
Ours (MD-REM++)	(10%,6%)	0.729	0.732	0.415	0.126
Ours (MD-REM++)	(10%;9%)	0.730	0.732	0.443	0.152

Table 2: Additional evaluations for Table 3 in the main paper. Here we report AP @ IoU 0.7 for these experiments.

Similar to the pedestrian experiment metrics, we further introduce a Top-5% rare subset metric. By mining more data via increasing the mining budget using our REM approach, we are able to further close up the gap with the fully labeled model with a much more reduced total labeling cost.

Bibliography

[1] Will Grathwohl, Ricky TQ Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. Ffjord: Free-form continuous dynamics for scalable reversible generative models. In *ICLR*, 2018.

[2] Charles R Qi, Yin Zhou, Mahyar Najibi, Pei Sun, Khoa Vo, Boyang Deng, and Dragomir Anguelov. Offboard 3d object detection from point cloud sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6134–6144, 2021.

[3] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2446–2454, 2020.

[4] Xinshuo Weng, Jianren Wang, David Held, and Kris Kitani. 3d multi-object tracking: A baseline and new evaluation metrics. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10359–10366. IEEE, 2020.

[5] Yin Zhou, Pei Sun, Yu Zhang, Dragomir Anguelov, Jiyang Gao, Tom Ouyang, James Guo, Jiquan Ngiam, and Vijay Vasudevan. End-to-end multi-view fusion for 3d object detection in lidar point clouds. In *Conference on Robot Learning*, pages 923–932. PMLR, 2020.