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

Abstract. Continued improvements in deep learning architectures have steadily advanced the overall performance of 3D object detectors to levels on par with humans for certain tasks and datasets, where the overall performance is mostly driven by common examples. However, even the best performing models suffer from the most naive mistakes when it comes to rare examples that do not appear frequently in the training data, such as vehicles with irregular geometries. Most studies in the long-tail literature focus on class-imbalanced classification problems with known imbalanced label counts per class, but they are not directly applicable to the intra-class long-tail examples in problems with large intra-class variations such as 3D object detection, where instances with the same class label can have drastically varied properties such as shapes and sizes. Other works propose to mitigate this problem using active learning based on the criteria of uncertainty, difficulty, or diversity. In this study, we identify a new conceptual dimension - rareness - to mine new data for improving the long-tail performance of models. We show that rareness, as opposed to difficulty, is the key to data-centric improvements for 3D detectors, since rareness is the result of a lack in data support while difficulty is related to the fundamental ambiguity in the problem. We propose a general and effective method to identify the rareness of objects based on density estimation in the feature space using flow models, and propose a principled cost-aware formulation for mining rare object tracks, which improves overall model performance, but more importantly - significantly improves the performance for rare objects (by 30.97%).

Keywords: Intra-class Long Tail, Rare Example, Active Learning

1 Introduction

Long-tail learning is a challenging yet important topic in applied machine learning, particularly for safety-critical applications such as autonomous driving or medical diagnostics. However, even though imbalanced classification problems have been heavily studied in the literature, we have limited tools in defining, identifying, and improving on intra-class rare instances, such as irregularly shaped vehicles or pedestrians in Halloween costumes, since they come from a diverse open set of anything but common objects. Inspired by Leo Tolstoy's famous

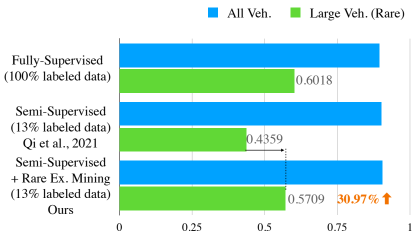


Fig. 1: Vehicle 3D object detection Average Precision (AP) on the Waymo Open Dataset with fully-/semi-supervised learning. While standard semi-supervised learning (with a strong auto labeling teacher model [40]) can achieve on par results with fully supervised method on the common cases, the performance gap on rare objects (e.g. large vehicles) is significant (60.18 v.s. 43.59). Our method is able to close this gap using rare example mining.

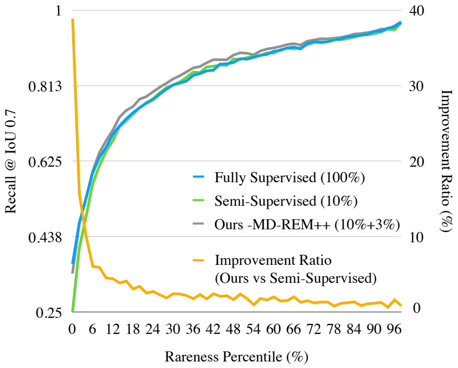


Fig. 2: Correlation between inferred rareness percentile (lower is more rare) and model performance for subsets of ground truth, indicated by recall. In all models (from fully-supervised to semi-supervised), model performance is strongly correlated to the rareness measure obtained from the log probability inferred by the flow model. By mining a mere 3% of remaining data, our model significantly improves upon the semi-supervised detector, with big gains in the rare intra-class long-tail.

quote, we observe: “Common objects are all alike; Every rare object is rare in its own way”.

We refer to the spectrum of such rare instances as the *intra-class long-tail*, where we do not have the luxury of prespecified class-frequency-based rareness measurements. Objects of the intra-class long-tail can be of particular importance in 3D detection due to its safety relevance. While overall performance for modern 3D detectors can be quite high, we note that even fully supervised models perform significantly worse on rare subsets of the data, such as large vehicles (Fig. 1). The problem is exacerbated by semi-supervised learning, a popular and cost-efficient approach to quickly scale models on larger datasets where average model performance have been shown to be on par with fully-supervised counterparts using a fraction of the labeled data.

Several challenges make it difficult for targeted improvement on the intra-class long-tail for 3D detection. First, as box regression is an important aspect of object detection, conventional long-tail learning approaches utilizing class frequencies, or active learning approaches utilizing entropy or margin uncertainties that depend on classification output distributions are not applicable. Second, since labeling cost given a run segment is proportional to the number of la-

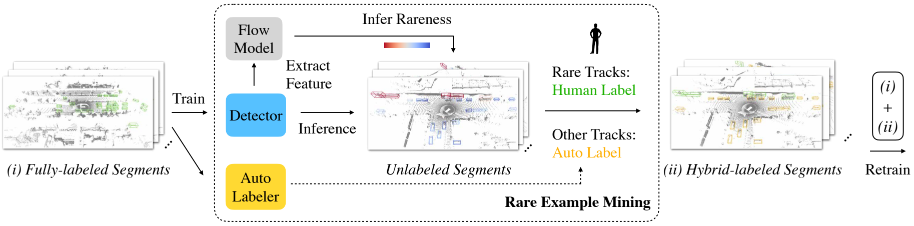


Fig. 3: Overview of the Rare Example Mining (REM) pipeline. Our detector, bootstrap-trained on a smaller pool of fully labeled segments, extracts features for a flow model to infer the log probability of every detected instance, which is a strong indicator of rareness. The rare tracks in the unlabeled segments are sent for human labeling while all remaining tracks are labeled using an offboard auto-labeler. The combined datasets is then used for retraining the detector, resulting in an overall performance boost, particularly on rare examples.

beled instance tracks, not frames, we require a more granular mining approach that gracefully handles missing labels for objects in the scene. Last but not least, unlike long-tail problems for imbalanced classification tasks, it is challenging to define which examples belongs to the intra-class long-tail, which leads to difficulty in evaluating and mining additional data to improve the long-tail performance of these models.

In light of these challenges, we propose a generalizable yet effective way to measure and define rareness as the density of instances in the latent feature space. We discover that normalizing flow models are highly effective for feature density estimation and robust for anomaly detection, contrary to negative results on anomaly detection using normalizing flows directly on high dimensional image inputs, as reported by prior work [38]. We present a cost-aware formulation for track-level data-mining and active learning using the rareness criteria, as 3D object labeling cost is often proportional to the number of unique tracks in each run segment. We do this in conjunction with a powerful offboard 3D auto-labeler [40, 58] for filling in missing data, and show stronger model improvement compared to difficulty, uncertainty, or heuristics based active learning baselines, particularly for objects in the tail distributions.

Furthermore, we investigate rareness as a novel data-mining criterion, in relation to the conventional uncertainty or error-based mining methods. Though models tend to perform poorly on either rare or hard examples, we note a clear distinction between the concept of rare versus hard. In this discussion, “rare” maps to epistemic uncertainty (reducible error) where the model is uncertain due to a lack of data support in the training set, while “hard” maps to aleatoric uncertainty (irreducible error), where the model is uncertain due to the fundamental ambiguity and uncertainty of the given problem, for example, if the target object is heavily occluded. We further illustrate that while conventional uncertainty estimates (such as ensembling methods) will uncover both hard and rare objects, filtering out hard examples will result in a significantly higher concentration of rare examples which significantly improves active learning performance, underscoring the importance of rare examples in active learning.

In summary, the main contributions of this work are:

- We identify rareness as a novel criterion for data mining and active learning, for improving model performance for problems with large intra-class variations such as 3D detection.
- We propose an effective way of identifying rare objects by estimating latent feature densities using a flow model, and demonstrate a strong correlation between estimated log probabilities, known rare subcategories, and model performance.
- We propose a fine-grained, cost-aware, track level mining methodology for 3D detection that utilizes a powerful offboard 3D auto-labeler for annotating unlabeled objects in partially labeled frames, resulting in a strong performance boost (30.97%) on intra-class long-tail subcategories compared to convetional semi-supervised baselines.

2 Related Work

Long-tail visual recognition: Long-tail is conventionally defined as an imbalance in a multinomial distribution between various different class labels, either in the image classification context [8, 24, 26, 27, 36, 55, 62, 64], dense segmentation problems [20, 23, 52, 53, 56, 59], or between foreground / background labels in object detection problems [33, 34, 50, 51, 60]. Existing approaches for addressing class-imbalanced problems include resampling (oversampling tail classes or head classes), reweighitng (using inverse class frequency, effective number of samples [8]), novel loss function design [1, 34, 50–52, 63], meta learning for head-to-tail knowlege transfer [7, 27, 35, 55], distillation [32, 57] and mixture of experts [54].

However, there is little work targeting improvements for the intra-class long-tail in datasets with inherently large intra-class variations, or for regression problems. Zhu et al. [66] studies the long-tail problem for subcategories, but assumes given subcategory labels. Dong et al. [12] studies imbalance between fine-grained attribute labels in clothing or facial datasets. To the best of our knowledge, our work is among the first to address the intra-class long-tail in 3D object detection.

Active learning: In this work we mainly address pool-based active learning [45], where we assume an existing smaller pool of fully-labeled data along with a larger pool of unlabeled data, from which we actively select samples for human labeling. Existing active learning methods mainly fall under two categories, uncertainty-based and diversity-based methods. Uncertainty-based methods select new labeling targets based on criteria such as ensemble variance [2] or classification output distribution such as entropy, margin or confidence [6, 14, 21, 22, 25, 41] in the case of classification outputs. More similar to our approach are diversity-based approaches, that aim at balancing the distribution of training data while mining from the unlabeled pool [18, 19, 39, 44]. Gudovskiy et al. [18] further targets unbalanced datasets. However, these methods are developed for classification problems and are not directly applicable to the intra-class long-tail for detection tasks. Similar to our approach, Sinha et al. [47] proposes to learn data distributions in the latent space, though they employ a discriminator in a variational setting that does not directly estimate the density of each data sample. Segal et al. [43] investigated fine-grained active learning in the context self-driving vehicles using region-based selection with a focus on joint perception and prediction. Similar to our approach, Elezi et al. [13] uses auto-labeling to improve active learning performances for 2D detection tasks.

Flow models: Normalizing flow models are a class of generative models that can approximate probability distributions and efficiently and exactly estimate the densities of high dimensional data [4, 10, 11, 17, 28, 30, 42]. Various studies have reported unsuccessful attempts at using density estimations estimated by normalizing flows for detecting out-of-distribution data by directly learning to map from the high dimensional pixel space of images to the latent space of flow models [5, 38, 61], assigning higher probability to out-of-distribution data. However, similar to our finding, Kirichenko et al. [29] find that the issue can be easily mitigated by training a flow model on the features extracted by a pretrained model such as an EfficientNet pretrained on ImageNet [9], rather than directly learning on the input pixel space. This allows the model to better measure density in a semantically relevant space. We are among the first to use densities estimated by normalizing flows for identifying long-tail examples.

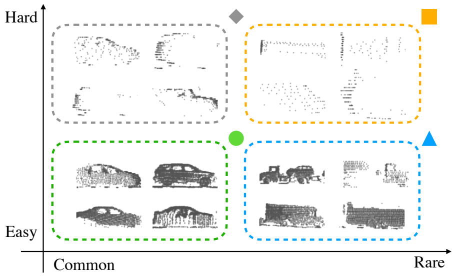
3 Methods

In this section, we present a general and effective method for mining rare examples based on density estimations from the data, which we refer to as data-centric rare example mining (REM). To offer further insights to rareness in relation to difficulty, we propose another conceptually simple yet effective method for mining rare examples by simply filtering out hard examples from overall uncertain examples. In Section 4.2, we show that combining both approaches can further improve long-tail performance. Last but not least, we propose a cost-aware, fine-grained track-level active learning method that aggregates per-track rareness as a selection criteria for requesting human annotation, and utilize a powerful off-board 3D auto-labeler for populating unmined, unlabeled tracks to maximize the utility of all data when retraining the model.

3.1 Rare Example Mining

Data-centric Rare Example Mining (D-REM) The main intuition behind data-centric REM is that we measure the density of every sample in a learned feature embedding space as an indicator for rareness.

The full data-centric REM workflow (see Fig. 3) consists of the following steps. First, we pretrain the detection model on an existing source pool of fully-labeled data that might be underrepresenting long-tail examples. Second, we



Mining Method	Mined Data	
Random Mining	◆ + ■ + ▲ + ●	
Hard Example Mining	◆ + ■	
Ensemble Mining	◆ + ■ + ▲	
Data-centric Rare Example	■ + ▲	*Ours*
Model-centric Rare Example	▲	
M+D Rare Example Mining	▲	

Fig. 4: Hard (aleatoric uncertainty) is a fundamentally different dimension compared to rare (epistemic uncertainty). Our REM method directly targets rare subsets of the data. Our Data-centric REM method directly estimates rareness based on inferred probabilities by a normalizing flow model trained on learned feature vectors, while our Model-centric REM method performs hard example filtering on top of generic uncertain objects mined by the ensemble mining approach. We further combine the two approaches (MD-REM) by performing hard example filtering on top of D-REM to increase easy-rare examples.

use the pretrained task model to run inference over the source pool along with a large unlabeled pool of data, and extract per-instance raw feature vectors via Region-of-interest (ROI) pooling, followed by PCA dimensionality reduction and normalization. We then train a normalizing flow model over the feature vectors to estimate per-instance rareness (negative log probability) for data mining.

Object Feature Extraction: As previously mentioned, one major difference between our proposed approach for estimating rare examples, compared with earlier works in the literature that were not successful in using normalizing flow for out-of-distribution detection [5, 38, 61], is that we propose to estimate the probability density of each example in the latent feature space of pretrained models to leverage the semantic similarity between objects for distinguishing rare instances. As observed by Kirichenko et al. [29], normalizing flow directly trained on high dimensional raw input features tend to focus more on local pixel correlations rather than semantics as it doesn't leverage high-level embeddings.

We extract per-object feature embeddings from the final Birds-Eye-View (BEV) feature map of a 3D object detector via region of interest (ROI) max-pooling [16] by cropping the feature map with the prediction boxes. We mainly apply this for our implementation of the state-of-the-art MVF [40, 65] 3D detector, though the process is generally applicable to majority of detectors that produce intermediate feature maps [31, 37, 49].

We further perform principal component analysis (PCA) for dimensionality reduction for improved computational efficiency, followed by normalization on the set of raw feature vectors $X_{\text{roi}} \in \mathbb{R}^{n \times d}$ obtained via ROI pooling

$$X_{\text{pca}} = (X_{\text{roi}} - mean(X_{\text{roi}}))W_{\text{pca}}^T \tag{1}$$

$$X_{\text{norm}} = X_{\text{pca}} \ / \ std(X_{\text{pca}}) \tag{2}$$

where $W_{\text{pca}} \in \mathbb{R}^{k\times d}$ is a weight matrix consisting of the top-k PCA components, $\text{mean}(\cdot) : \mathbb{R}^{n\times d} \mapsto \mathbb{R}^d$, $std(\cdot) : \mathbb{R}^{n\times d} \mapsto \mathbb{R}^d$ are the mean and standard deviation operators along the first dimension.

In summary, the training dataset for our flow model consists of normalized feature vectors after PCA-transformation obtained via ROI max-pooling the final feature map of 3D detectors using predicted bounding boxes.

$$\mathcal{D}_x = \{X_{\text{norm}}[i], \forall i \in [0, n)\} \tag{3}$$

Rareness Estimation Using Normalizing Flow: We use the continuous normalizing flow models for directly estimating the log probability of each example represented as a feature vector $\boldsymbol{x}$. We present a quick review of normalizing flows below.

Typical normalizing flow models [28] consist of two main components: a base distribution $p(\boldsymbol{z})$, and a learned invertible function $f_{\boldsymbol{\theta}}(\boldsymbol{x})$, also known as a bijector, where $\boldsymbol{\theta}$ are the learnable parameters of the bijector, $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ is the forward method and $f_{\boldsymbol{\theta}}^{-1}(\boldsymbol{x})$ is the inverse method. The base distribution is generally chosen to be an analytically tractable distribution whose probability density function (PDF) can be easily computed, such as a spherical multivariate Gaussian distribution, where $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; 0, \boldsymbol{I})$. A learnable bijector function can take many forms, popular choices include masked scale and shift functions such as RealNVP [11, 28] or continuous bijectors utilizing learned ordinary differential equation (ODE) dynamics [4, 17].

The use of normalizing flows as generative models has been heavily studied in the literature [28], where new in-distribution samples can be generated via passing a randomly sampled latent vector through the forward bijector:

$$\boldsymbol{x} = f_{\boldsymbol{\theta}}(\boldsymbol{z}), \quad \text{where } \boldsymbol{z} \sim p(z) \tag{4}$$

However, in this work, we are more interested in using normalizing flows for estimating the exact probabilities of each data example. The latent variable corresponding to a data example can be inferred via $\boldsymbol{z} = f_{\boldsymbol{\theta}}(\boldsymbol{x})$. Under a change-of-variables formula, the log probability of a data sample can be estimated as:

$$\log p_\theta(\boldsymbol{x}) = \log p(f_{\boldsymbol{\theta}}(\boldsymbol{x})) + \log|\det(df_{\boldsymbol{\theta}}(\boldsymbol{x})/d\boldsymbol{x})| \tag{5}$$

$$= \log p(\boldsymbol{z}) + \log|\det(d\boldsymbol{z}/d\boldsymbol{x})| \tag{6}$$

The first term, $\log p(\boldsymbol{z})$, can be efficiently computed from the PDF of the base distribution, whereas the computation of the log determinant of the Jacobian: $\log|\det(df_{\boldsymbol{\theta}}(\boldsymbol{x})/d\boldsymbol{x})|$ vary based on the bijector type.

The training process can be described as a simple maximization of the expected log probability of the data (or equivalently minimization of the expected negative log likelihood of the parameters) from the training data $\mathcal{D}_{\boldsymbol{x}}$ and can be learned via batch stochastic gradient descent:

$$\arg\min_{\theta} \ \mathbb{E}_{x\sim\mathcal{D}_{\boldsymbol{x}}}[-\log p_{\boldsymbol{\theta}}(\boldsymbol{x})] \tag{7}$$

In our experiments, we choose the base distribution $p(\boldsymbol{z})$ to be a spherical multivariate Gaussian $\mathcal{N}(\boldsymbol{z}; 0, \boldsymbol{I})$, and we use the FFJORD [17] bijector.

For the final rare example scoring function for the i-th object, r_i, we have:

$$r_i = -\log p_{\boldsymbol{\theta}}(\boldsymbol{x_i}) \tag{8}$$

Model-centric Rare Example Mining (M-REM) We present an alternative model-centric formulation for REM that is conceptually easy and effective, yet illustrative of the dichotomy between rare and hard examples. Different from the data-centric REM perspective, model-centric REM leverages the divergence among an ensemble of detectors as a measurement of total uncertainty.

Different from methods that directly use ensemble divergence as a mining critera for active learning [2], our key insight is that while ensemble divergence is a good measurement of the overall uncertainties for an instance, it could be either due to the problem being fundamentally difficult and ambiguous (i.e., hard), or due to the problem being uncommon and lack training support for the model (i.e., rare). In the case of 3D object detection, a leading reason for an object being physically hard to detect is occlusion and low number of LiDAR points from the object. Conceptually, adding more hard examples such as faraway and heavily occluded objects with very few visible LiDAR points would not be helpful, as these cases are fundamentally ambiguous and cannot be improved upon simply with increased data support.

A simple approach for obtaining rare examples, hence, is to filter out hard examples from the set of overall uncertain examples. In practice, a simple combination of two filters: (i) low number of LiDAR points per detection example, or (ii) a large distance between the detection example and the LiDAR source, proves to be surprisingly effective for improving model performance through data mining and active learning.

We implement model-centric REM as follows. Let $\mathcal{M} = \{M_1, M_2, \cdots, M_N\}$ be a set of N independently trained detectors with identical architecture and training configurations, but different model initialization. Denote detection score for the i-th object by the j-th detector as s_i^j. s_i^j is set to zero if there is a missed detection. The detection variance for the i-th object by the model ensemble $\mathcal{M}$ is defined as:

$$v_i = \frac{1}{N}\sum_{j=1}^{N}(s_i^j - \frac{1}{N}\sum_{k=1}^{N} s_i^k)^2 \tag{9}$$

For hard example filtering, denote the number of LiDAR points within the i-th object as p_i, and the distance of the i-th object from the LiDAR source as d_i. A simple hard example filter function can be defined as:

$$h_i = 1 \text{ if } (p_i > \tilde{p}) \;\&\; (d_i < \tilde{d}) \text{ else } 0 \tag{10}$$

where $\tilde{p}, \tilde{d}$ are the respective point threshold and distance thresholds. In our experiments, we have $N = 5, \tilde{p} = 200, \tilde{d} = 50$ (meters).

The final rare example scoring function for the i-th object, r_i, can be given as:

$$r_i = h_i * v_i \tag{11}$$

3.2 Track-level REM for Active Learning

To apply our REM method towards active learning as a principled way of collecting rare instances from a large unlabeled pool in a cost-effective manner, we propose a novel track-level mining and targeted annotation strategy in conjunction with a high-performance offboard 3D auto-labeler for infilling missing labels. We choose to mine at the track-level because labeling tools are optimized to label entire object tracks, which is cheaper than labeling per frame. Please refer to Fig. 3 for an overview of the active learning pipeline and Algorithm 1 for a detailed breakdown of the mining process.

First, starting with a labeling budget of K tracks, we score each detected object from the unlabeled dataset using one of the rare example scoring functions above (Eq. (8, 11)). Starting from the detection object with the highest rareness score, we sequentially route each example to human labelers for labeling the entire track T corresponding to the object and add the track to the set of mined and human-labeled tracks $\mathcal{S}_h$. Then all model detections that intersect with T (> 0 IoU) are removed. This procedure is iteratively performed until the number of tracks in $\mathcal{S}_h$ reaches the budget of K. All auto-labeled tracks $\mathcal{S}_a$ that intersect with $\mathcal{S}_h$ are removed, and the two sets of tracks are merged into a hybrid, fully-labeled dataset $\mathcal{S} = \mathcal{S}_a \cup \mathcal{S}_h$.

4 Experiments

We use the Waymo Open Dataset [48] as the main dataset for our investigations due to its unparalleled diversity based on geographical coverage, compared with other camera+LiDAR datasets available [3, 15], as well as its large industry-level scale. The Waymo Open Dataset consists of 1150 scenes that span 20 seconds, recorded across a range of weather conditions in multiple cities.

In the experiments below, we seek to answer three questions: (1) Does model performance correlate with our rareness measurement for intra-class long-tail (Section 4.1), (2) Can our proposed rare example mining methodology successfully find and retrieve more rare examples (Section 4.1), and (3) Does adding rare data to our existing training data in an active learning setting improve overall model performance, in particular for the long-tail (Section 4.2).

4.1 Rare Example Mining Analysis

In this section, we investigate the ability of the normalizing flow model in our data-centric REM method for detecting intra-class long-tail examples.

Correlation: Rareness and Performance: We investigate the correlation between the rareness metric (as indicated by low inferred log probability score on ground-truth labels), and the associated model performance on these examples, as measured by recall on GT examples grouped by rareness. We present the results in Fig. 2. All ground-truth examples are grouped by sorting along their inferred log probability (from an MVF and flow model trained on 100% data)

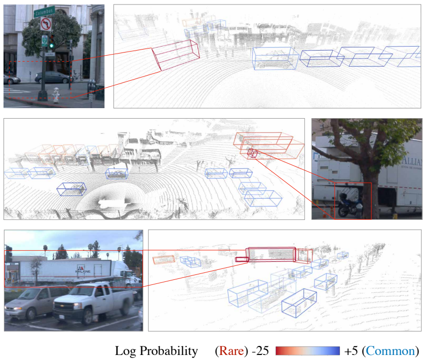


Fig. 5: Visualization of the rarest object tracks in the Waymo Open Dataset based on log probability inferred by the data-centric REM algorithm, where low log probability indicates rareness. The most rare instances include incorrectly labeled ground truth boxes, motorcycle underneath a trailer, and large vehicles.

```
Input
    Model detections $\mathcal{D}_b = b_1, b_2, \cdots, b_n$ (sort by descending rareness score)
    Auto-labeled tracks $\mathcal{S}_a = \{T'_1, T'_2, \cdots, T'_m\}$
    Labeling budget $K$
Output
    Fully labeled tracks $\mathcal{S} = \mathcal{S}_h \cup (\mathcal{S}_a - (\mathcal{S}_h \cap \mathcal{S}_a))$
procedure TrackMining
    $\mathcal{S}_h \leftarrow \{\emptyset\}$
    while $|\mathcal{S}_h| < K$ do
        $b \leftarrow \mathcal{D}_b.pop(0)$
        if HumanCheckExists($b$) then
            $T$ =HumanLabelTrackFromBox($b$)
            $\mathcal{S}_h.push(T)$
            $\mathcal{D}_b \leftarrow$ DiscardIntersectingBoxes($\mathcal{D}_b, T$)
        end if
    end while
    $\mathcal{S}_a \leftarrow \mathcal{S}_a - (\mathcal{S}_h \cap \mathcal{S}_a)$
return $\mathcal{S} = \mathcal{S}_a \cup \mathcal{S}_h$
end procedure
```

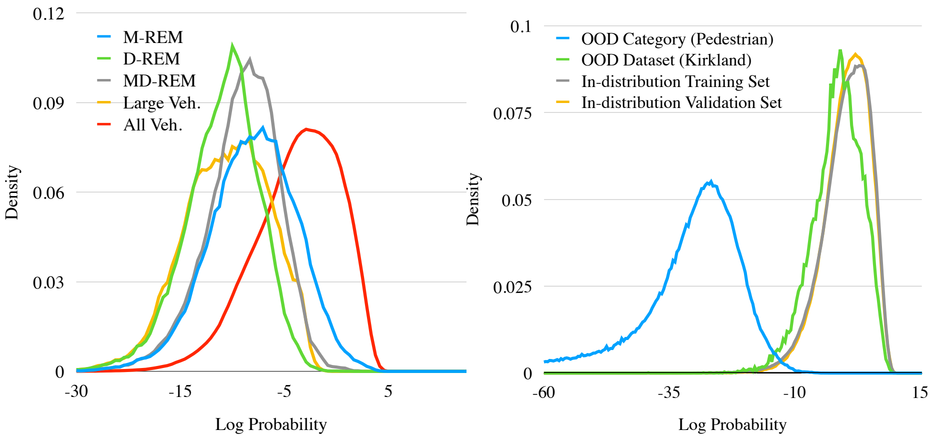
Algorithm 1: Track-level REM

into 2% bins. Recall metric for different experiments are computed for each bin. More details on our active learning experiment will be presented in Section 4.2.

We derive two main conclusions: (1) the performance for all models are strongly correlated with our proposed rareness measurement, indicating our flow probability-based estimation of rareness is highly effective. (2) Our proposed rare example mining method achieves significant performance improvement on rare examples compared to the original semi-supervised baseline using a small fraction of additional human-labeled data.

Visualizing Rare Examples: We visualize the rarest ground-truth examples from the Waymo Open Dataset as determined by the estimated log probability of every instance. We aggregate the rareness score for every track by taking the mean log probability of the objects from different frames in each track. We then rank the objects by descending average log probability. See Fig. 5 for a visualization of the rarest objects in the dataset.

The rarest ground-truth objects include boxes around vehicle parts (protruding ducts, truck loading ramp) and oversized or irregularly shaped vehicles (trucks, flatbed trailers), which match our intuition regarding rare vehicles. Moreover we discover a small number of mislabeled ground-truth instances among the rarest examples. This illustrates that rare example detection is an out-



(a) Different vehicle subsets.

(b) In-domain vs OOD distributions.

Fig. 6: Distributional sensitivity of the flow model trained on the vehicle class of the Waymo Open Dataset [48]. (a) Log probability distribution of different vehicle subsets (size subsets and REM mined subsets). (b) Log probability distribution of in-/out-of-distribution examples.

of-distribution detection problem. Intra-class long-tail examples, in one sense, can be defined as in-category, out-of-distribution examples.

Distributional Sensitivity of the Flow Model: In-light of the observation that rare example mining is inherently an out-of-distribution detection problem, we seek to perform a more quantitative analysis of the model's sensitivity to out-of-distribution instances. See Fig. 6 for a detailed breakdown of the analysis. In Fig. 6a, we compare the flow model's inferred log probability distributions between vehicle boxes of different sizes. Vehicle size is defined as the max between box length, width and height. We perform a simple partition for all vehicle examples along the size dimension: regular vehicles as size $\in [3, 7)$ (m), and large vehicles as size $\in [7, \infty)$ (m). Our flow model assigns significantly higher overall log probability for the subset of regular-sized vehicles (96.18% of total), compared to rare subsets such as large vehicles. Note that we leverage vehicle size as an sanity check for the general REM method to distinguish between known rare and common distribution.

Furthermore, we validate that the flow model is effective at detecting out-of-distribution examples (Fig. 6b). The flow model infers almost-identical log probability distributions between the training and validation sets, while assigning lower probabilities to vehicles from an out-of-distribution set (the Kirkland set from the Waymo Open Dataset, collected from a different geographical region with mostly rainy weather condition). Moreover, the model assigns significantly lower probabilities on OOD categories (Pedestrian) if we perform ROI pooling using the pedestrian ground-truth boxes to extract pedestrian feature vectors

Mining Criteria	Ratio of Large
Random Uniform	2.60%
Ensemble[2]	13.72%
Model-centric REM	24.61%
Data-centric REM	30.60%
Model+Data-centric REM	**31.86**%

Table 1: Composition of mined tracks. We use the ratio of large (> 7m) objects as a reference for measuring the ratio of rare tracks mined by different approaches. REM is able to mine a higher proportion of rare instances.

Experiment	Human Labels	All	Regular	Large
Fully Supervised	100%;0%	0.895	0.900	0.602
Ours	10%;3%	0.904	**0.904**	0.571
Ours	10%;6%	0.904	0.903	**0.612**
Our	10%;9%	**0.905**	**0.904**	0.606

Table 2: Impact of mining budget on model performance. With a small increase in mining budget (6%), we (MD-REM++) can match the performance of a fully-supervised model on both ends of the spectrum.

from the vehicle model and query the log probability distribution against the flow model.

4.2 Rare Example Mining for Active Learning

To demonstrate the applicability of the REM approach for targeted improvement of the model's performance in the intra-class long-tail, we utilize track-level REM for active learning, as detailed in Section 3.2.

Experiment Setup: Our experiment setup is as follows. Following Qi et al. [40], we perform a random split on the main training set of the Waymo Open Dataset [48] into a 10% fully-labeled source pool, and a remaining 90% as a larger "unlabeled" pool, from which we withhold ground-truth labels. We first train our main model on the fully-labeled source pool, and perform track-level data mining on the remaining unlabeled pool using various methods, including our proposed data-centric and model-centric REM approaches. For all active learning baseline experiments, we mine for a fixed budget of 1268 tracks, amounting to $\sim 3\%$ of all remaining tracks.

Our main model consists of a single-frame MVF detector [40, 65]. While in all baseline experiments we utilize the main model for self-labeling unlabeled tracks in the unlabeled pool, we demonstrate that using a strong offboard 3D auto-labeler [40] trained on the same existing data can further boost the overall performance of our REM approach.

Composition of Mined Tracks: We first analyze the composition of the mined tracks, in all cases 1268 tracks obtained using various mining approaches (see Table 1).

We derive three main findings from the composition analysis: (1) Data-centric REM is able to effectively retrieve known rare subsets, upsampling large objects by as much as 1214%. (2) Comparing model-centric REM to ensemble mining method, a simple hard example filtering operator leads to drastically upsampled

(a) Reference experiments w/o active learning.

Experiment	Human Labels	All	Regular	Large
Partial-supervised	10%;0%	0.845	0.853	0.378
Semi-supervised (SL)	10%;0%	0.854	0.864	0.350
Semi-supervised (AL)	10%;0%	0.902	0.910	0.419
Fully-supervised	100%;0%	0.895	0.900	0.602

(b) Oracle active learning experiments.

Experiment	Human Labels	All	Regular	Large
Oracle Hard [46]	10%;3%	0.865	0.875	0.341
Oracle Size	10%;3%	0.869	0.875	0.583

(c) Main active learning experiments.

Experiment	Human Labels	All	Regular	Large
Partial-supervised	10%;0%	0.845	0.853	0.378
Random	10%;3%	0.873	0.881	0.355
Predict Size	10%;3%	0.865	0.871	0.498
Ensemble [2]	10%;3%	0.869	0.879	0.353
Ours (M-REM)	10%;3%	0.886	0.893	0.478
Ours (D-REM)	10%;3%	0.882	0.888	0.483
Ours (D-REM++)	10%;3%	**0.906**	**0.913**	0.533
Ours (MD-REM++)	10%;3%	0.904	0.909	**0.571**

Table 3: Active learning experiment results. Our method significantly improves model performance across the spectrum, particularly significantly on rare subsets. We denote human label ratio as $(\%s, \%t)$ to indicate the model being trained with $\%s$ of full-labeled run segments, along with $\%t$ of the remaining tracks that is mined and labeled.

rare instances, signifying the dichotomy of rare and hard. By using a hard example filter we can significantly increase the ratio of rare examples among mined tracks. (3) Combining model and data-centric REM (by further performing hard example filtering from instanced mined by data-centric REM) further boosts the ratio of large vehicles.

Active Learning Experiment: We present our active learning experiments in Table 3. Results are on vehicles from the Waymo Open Dataset [48], reported as AP at IoU 0.5. We compute subset metrics on all vehicles ("All"), regular vehicles ("Regular") of size within $3 - 7$m, large vehicles ("Large") of size > 7 m. "Regular" subset is a proxy of the common vehicles, while "Large" subset is a proxy for rare.

In Table 3a, we present performances of the single-frame MVF model trained on different compositions of the data. We denote semi-supervised method using self-labeled segments as "SL" and auto-labeled segments as "AL". The main observation is that although auto-labeling can significantly improve overall model performance, in particular for common (regular-sized) vehicles, the resulting model performance is significantly weaker for rare subsets, motivating our REM approach.

For the active learning experiments, we first compare two oracle-based approaches (Table 3b) that utilize 100% ground-truth knowledge for the mining process. "Oracle Hard" is an error-driven mining method inspired by [46], that ranks tracks by s = IoU(GT, Pred) * Probability_Score(Pred) to mine tracks which either the base model made a wrong prediction on, or made an inconfident prediction. "Oracle Size" explicitly mines 3% of ground-truth tracks whose box size is $> 7m$. The main observation is that error-based mining favors difficult examples which do not help improve model performance. Though size-based mining can effectively improve large vehicle performance, it solely improves large vehicles and does not help on regular vehicles.

We then compare across a suite of active learning baselines and our proposed REM methods (Table 3c). "Random" mines the tracks via randomized selection, "Predict Size" mines tracks associated with the largest predicted boxes, and "Ensemble" mines the tracks with highest ensemble variance (Eq. (9)). For our proposed REM methods, we prefix model-centric REM approaches with "M-", data-centric approaches with "D-", and a hybrid approach leveraging hard-example filtering on top of data-centric approaches with "MD-". To further illustrate the importance of a strong offboard auto-labeler, we add auto-labeler to our method, denoting the experiments with "++".

The active learning experiments show that: (1) Both data-centric and model-centric approaches significantly help to improve performance on the rare subset, and a combination of the two can further boost the long-tail performance, (2) While heuristics based mining methods ("Predict Size") can achieved targeted improvement for large vehicles, it likely fails to capture other degrees of rareness, resulting in lower overall performance.

5 Ablation studies

We further study the impact of increasing mining budget on our REM approach (Table 2). With a small increase of mining budget (6%), we can match the performance of a fully-supervised model for both common and rare subsets.

6 Discussions and Future Work

In this work, we illustrate the limitations of learned detectors with respect to rare examples in problems with large intra-class variations, such as 3D detection. We propose an active learning approach based on data-centric and model-centric rare example mining which is effective at discovering rare objects in unlabled data. Our active learning approach, combined with a state-of-the-art semi-supervised method can achieve full parity with a fully-supervised model on both common and rare examples, utilizing as little as 16% of human labels.

A limitation of this study is the scale of the existing datasets for active learning, where data mining beyond the scale of available datasets is limited. Results on a larger dataset will be more informative. Our work shares the same risks and opportunities for the society as other works in 3D detection.

Future work includes extending the REM approach beyond 3D detection, including other topics in self-driving such as trajectory prediction and planning.

Acknowledgements We thank Marshall Tappen, Zhao Chen, Tim Yang, Abhishek Sinh and Luna Yue Huang for helpful discussions, Mingxing Tan for proofreading and constructive feedback, and anonymous reviews for in-depth discussions and feedback.

Bibliography

[1] Abdelkarim, S., Achlioptas, P., Huang, J., Li, B., Church, K., Elhoseiny, M.: Long-tail visual relationship recognition with a visiolinguistic hubless loss (2020)

[2] Beluch, W.H., Genewein, T., Nürnberger, A., Köhler, J.M.: The power of ensembles for active learning in image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9368–9377 (2018)

[3] Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11621–11631 (2020)

[4] Chen, R.T., Rubanova, Y., Bettencourt, J., Duvenaud, D.: Neural ordinary differential equations. arXiv preprint arXiv:1806.07366 (2018)

[5] Choi, H., Jang, E., Alemi, A.A.: Waic, but why? generative ensembles for robust anomaly detection. arXiv preprint arXiv:1810.01392 (2018)

[6] Choi, J., Elezi, I., Lee, H.J., Farabet, C., Alvarez, J.M.: Active learning for deep object detection via probabilistic modeling. arXiv preprint arXiv:2103.16130 (2021)

[7] Chu, P., Bian, X., Liu, S., Ling, H.: Feature space augmentation for long-tailed data. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16, pp. 694–710, Springer (2020)

[8] Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9268–9277 (2019)

[9] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, pp. 248–255, Ieee (2009)

[10] Dinh, L., Krueger, D., Bengio, Y.: Nice: Non-linear independent components estimation. arXiv preprint arXiv:1410.8516 (2014)

[11] Dinh, L., Sohl-Dickstein, J., Bengio, S.: Density estimation using real nvp. arXiv preprint arXiv:1605.08803 (2016)

[12] Dong, Q., Gong, S., Zhu, X.: Class rectification hard mining for imbalanced deep learning. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1851–1860 (2017)

[13] Elezi, I., Yu, Z., Anandkumar, A., Leal-Taixe, L., Alvarez, J.M.: Not all labels are equal: Rationalizing the labeling costs for training object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14492–14501 (2022)

[14] Gal, Y., Islam, R., Ghahramani, Z.: Deep bayesian active learning with image data. In: International Conference on Machine Learning, pp. 1183–1192, PMLR (2017)

[15] Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The International Journal of Robotics Research **32**(11), 1231–1237 (2013)

[16] Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 580–587 (2014)

[17] Grathwohl, W., Chen, R.T., Bettencourt, J., Sutskever, I., Duvenaud, D.: Ffjord: Free-form continuous dynamics for scalable reversible generative models. In: ICLR (2018)

[18] Gudovskiy, D., Hodgkinson, A., Yamaguchi, T., Tsukizawa, S.: Deep active learning for biased datasets via fisher kernel self-supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9041–9049 (2020)

[19] Guo, Y.: Active instance sampling via matrix partition. In: NIPS, pp. 802–810 (2010)

[20] Gupta, A., Dollar, P., Girshick, R.: Lvis: A dataset for large vocabulary instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5356–5364 (2019)

[21] Harakeh, A., Smart, M., Waslander, S.L.: Bayesod: A bayesian approach for uncertainty estimation in deep object detectors. In: 2020 IEEE International Conference on Robotics and Automation (ICRA), pp. 87–93, IEEE (2020)

[22] Holub, A., Perona, P., Burl, M.C.: Entropy-based active learning for object recognition. In: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–8, IEEE (2008)

[23] Hsieh, T.I., Robb, E., Chen, H.T., Huang, J.B.: Droploss for long-tail instance segmentation. arXiv preprint arXiv:2104.06402 (2021)

[24] Jamal, M.A., Brown, M., Yang, M.H., Wang, L., Gong, B.: Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7610–7619 (2020)

[25] Joshi, A.J., Porikli, F., Papanikolopoulos, N.: Multi-class active learning for image classification. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2372–2379, IEEE (2009)

[26] Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., Kalantidis, Y.: Decoupling representation and classifier for long-tailed recognition. arXiv preprint arXiv:1910.09217 (2019)

[27] Kim, J., Jeong, J., Shin, J.: M2m: Imbalanced classification via major-to-minor translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13896–13905 (2020)

[28] Kingma, D.P., Dhariwal, P.: Glow: Generative flow with invertible 1x1 convolutions. arXiv preprint arXiv:1807.03039 (2018)

[29] Kirichenko, P., Izmailov, P., Wilson, A.G.: Why normalizing flows fail to detect out-of-distribution data. In: NIPS (2020)

[30] Kobyzev, I., Prince, S., Brubaker, M.: Normalizing flows: An introduction and review of current methods. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)

[31] Lang, A.H., Vora, S., Caesar, H., Zhou, L., Yang, J., Beijbom, O.: Pointpillars: Fast encoders for object detection from point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12697–12705 (2019)

[32] Li, T., Wang, L., Wu, G.: Self supervision to distillation for long-tailed visual recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 630–639 (2021)

[33] Li, Y., Wang, T., Kang, B., Tang, S., Wang, C., Li, J., Feng, J.: Overcoming classifier imbalance for long-tail object detection with balanced group softmax. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10991–11000 (2020)

[34] Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp. 2980–2988 (2017)

[35] Liu, B., Li, H., Kang, H., Hua, G., Vasconcelos, N.: Gistnet: a geometric structure transfer network for long-tailed recognition. arXiv preprint arXiv:2105.00131 (2021)

[36] Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., Yu, S.X.: Large-scale long-tailed recognition in an open world. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2537–2546 (2019)

[37] Meyer, G.P., Laddha, A., Kee, E., Vallespi-Gonzalez, C., Wellington, C.K.: Lasernet: An efficient probabilistic 3d object detector for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12677–12686 (2019)

[38] Nalisnick, E., Matsukawa, A., Teh, Y.W., Gorur, D., Lakshminarayanan, B.: Do deep generative models know what they don't know? In: International Conference on Learning Representations (2018), URL https://openreview.net/forum?id=H1xwNhCcYm

[39] Nguyen, H.T., Smeulders, A.: Active learning using pre-clustering. In: Proceedings of the twenty-first international conference on Machine learning, p. 79 (2004)

[40] Qi, C.R., Zhou, Y., Najibi, M., Sun, P., Vo, K., Deng, B., Anguelov, D.: Offboard 3d object detection from point cloud sequences. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6134–6144 (2021)

[41] Qi, G.J., Hua, X.S., Rui, Y., Tang, J., Zhang, H.J.: Two-dimensional active learning for image classification. In: 2008 IEEE conference on computer vision and pattern recognition, pp. 1–8, IEEE (2008)

[42] Rezende, D., Mohamed, S.: Variational inference with normalizing flows. In: International conference on machine learning, pp. 1530–1538, PMLR (2015)

[43] Segal, S., Kumar, N., Casas, S., Zeng, W., Ren, M., Wang, J., Urtasun, R.: Just label what you need: Fine-grained active selection for perception and prediction through partially labeled scenes. arXiv preprint arXiv:2104.03956 (2021)

[44] Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach (2017)

[45] Settles, B.: Active learning literature survey (2009)

[46] Shrivastava, A., Gupta, A., Girshick, R.: Training region-based object detectors with online hard example mining. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 761–769 (2016)

[47] Sinha, S., Ebrahimi, S., Darrell, T.: Variational adversarial active learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5972–5981 (2019)

[48] Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2446–2454 (2020)

[49] Sun, P., Wang, W., Chai, Y., Elsayed, G., Bewley, A., Zhang, X., Sminchisescu, C., Anguelov, D.: Rsn: Range sparse net for efficient, accurate lidar 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5725–5734 (2021)

[50] Tan, J., Lu, X., Zhang, G., Yin, C., Li, Q.: Equalization loss v2: A new gradient balance approach for long-tailed object detection. arXiv preprint arXiv:2012.08548 (2020)

[51] Tan, J., Wang, C., Li, B., Li, Q., Ouyang, W., Yin, C., Yan, J.: Equalization loss for long-tailed object recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11662–11671 (2020)

[52] Wang, J., Zhang, W., Zang, Y., Cao, Y., Pang, J., Gong, T., Chen, K., Liu, Z., Loy, C.C., Lin, D.: Seesaw loss for long-tailed instance segmentation. arXiv preprint arXiv:2008.10032 (2020)

[53] Wang, T., Li, Y., Kang, B., Li, J., Liew, J.H., Tang, S., Hoi, S., Feng, J.: Classification calibration for long-tail instance segmentation. arXiv preprint arXiv:1910.13081 (2019)

[54] Wang, X., Lian, L., Miao, Z., Liu, Z., Yu, S.X.: Long-tailed recognition by routing diverse distribution-aware experts. In: International Conference on Learning Representations (2021), URL https://openreview.net/forum?id=D9I3drBz4UC

[55] Wang, Y.X., Ramanan, D., Hebert, M.: Learning to model the tail. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 7032–7042 (2017)

[56] Wu, J., Song, L., Wang, T., Zhang, Q., Yuan, J.: Forest r-cnn: Large-vocabulary long-tailed object detection and instance segmentation. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 1570–1578 (2020)

[57] Xiang, L., Ding, G., Han, J.: Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In: European Conference on Computer Vision, pp. 247–263, Springer (2020)

[58] Yang, B., Bai, M., Liang, M., Zeng, W., Urtasun, R.: Auto4d: Learning to label 4d objects from sequential point clouds. arXiv preprint arXiv:2101.06586 (2021)

[59] Zang, Y., Huang, C., Loy, C.C.: Fasa: Feature augmentation and sampling adaptation for long-tailed instance segmentation. arXiv preprint arXiv:2102.12867 (2021)

[60] Zhang, C., Pan, T.Y., Li, Y., Hu, H., Xuan, D., Changpinyo, S., Gong, B., Chao, W.L.: A simple and effective use of object-centric images for long-tailed object detection. arXiv e-prints pp. arXiv–2102 (2021)

[61] Zhang, L., Goldstein, M., Ranganath, R.: Understanding failures in out-of-distribution detection with deep generative models. In: International Conference on Machine Learning, pp. 12427–12436, PMLR (2021)

[62] Zhao, Y., Chen, W., Tan, X., Huang, K., Xu, J., Wang, C., Zhu, J.: Improving long-tailed classification from instance level. arXiv preprint arXiv:2104.06094 (2021)

[63] Zheng, Y., Pal, D.K., Savvides, M.: Ring loss: Convex feature normalization for face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5089–5097 (2018)

[64] Zhong, Z., Cui, J., Liu, S., Jia, J.: Improving calibration for long-tailed recognition. arXiv preprint arXiv:2104.00466 (2021)

[65] Zhou, Y., Sun, P., Zhang, Y., Anguelov, D., Gao, J., Ouyang, T., Guo, J., Ngiam, J., Vasudevan, V.: End-to-end multi-view fusion for 3d object detection in lidar point clouds. In: Conference on Robot Learning, pp. 923–932, PMLR (2020)

[66] Zhu, X., Anguelov, D., Ramanan, D.: Capturing long-tail distributions of object subcategories. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 915–922 (2014)