Supplementary Materials
DatasetEquity: Are All Samples Created Equal?
In The Quest For Equity Within Datasets

A. Dataset cluster visualization

Camera images from various datasets are first passed through a feature extractor, followed by a low-dimensional feature projection (t-SNE). These low-dimensional (3D) features are then clustered together to reveal frames with similar semantics and perceptual details. Relative size of each of these clusters are then used as a proxy for quantifying likelihood of occurrence for each sample within those clusters. Figure S1 visualize these clusters in 2D for the four datasets analyzed in this work.
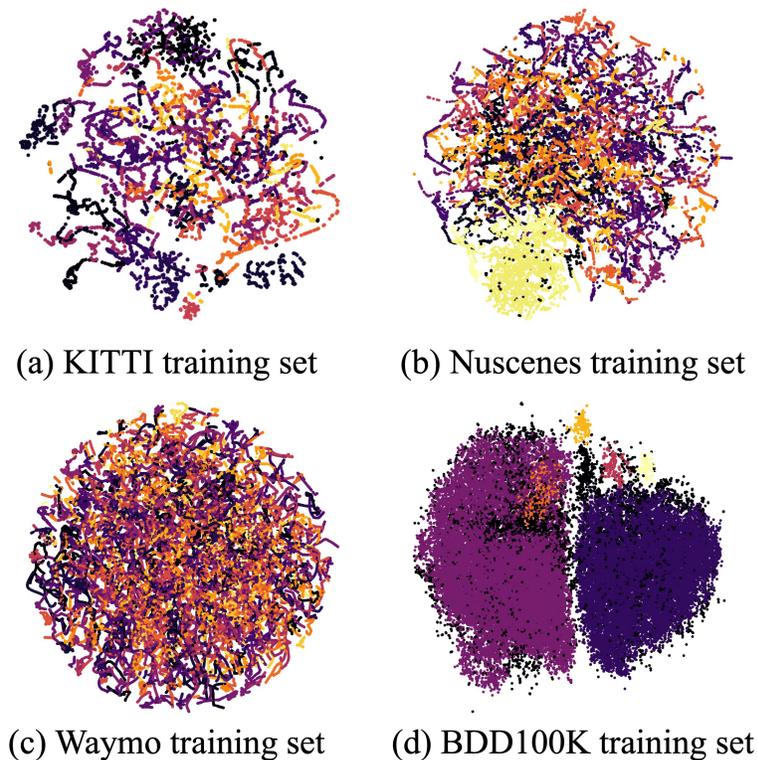

(a) KITTI training set (b) Nuscenes training set

(c) Waymo training set (d) BDD100K training set

Figure S1: NuScenes, KITTI, Waymo, and BDD100k dataset samples projected onto a 3-dimensional t-SNE space and then clustered using DBSCAN (only first 2-dimensions are visualized). Each color represents a unique cluster ID. Only a single front camera image was used to compute these clusters.

B. Dataset samples visualization

A few examples of samples from various clusters and their scaled likelihoods are shown in sections B.1, C, D, and E. These clusters were computed by first projecting 2048-dimensional image embeddings onto a lower 3-dimensional space using t-SNE, and then applying the DBSCAN algorithm. Likelihood of each sample is then computed as $\mathcal{L}_{s(i)} = \frac{|C_i|}{\max_i(|C_i|)}$, where C_i refers to the ith cluster, $|C_i|$ is the number of samples in i^{th} cluster, and $\mathcal{L}_{s(i)}$ is the scaled likelihood of samples corresponding to i^{th} cluster. Corresponding to each row in the visualized samples below, we also provide the measure for corresponding likelihood values. All images in a certain row are taken from the same cluster.

B.1. KITTI training dataset

(a) $\mathcal{L}_{s(i)} = 0.0192$

(b) $\mathcal{L}_{s(i)} = 0.0240$

(c) $\mathcal{L}_{s(i)} = 0.4855$

(d) $\mathcal{L}_{s(i)} = 0.6178$

(e) $\mathcal{L}_{s(i)} = 1.0$

Figure S2: Samples from various clusters in KITTI *training* dataset, and their likelihoods relative to the largest cluster within the dataset. Samples shown in the last row come from a cluster of size 416, and is the largest among all clusters.

C. nuScenes training dataset

(a) $\mathcal{L}_{s(i)} = 0.0034$

(b) $\mathcal{L}_{s(i)} = 0.0034$

(c) $\mathcal{L}_{s(i)} = 0.0038$

(d) $\mathcal{L}_{s(i)} = 0.3592$

(e) $\mathcal{L}_{s(i)} = 1.0$

Figure S3: Samples from various clusters in nuScenes *training* dataset, and their likelihoods relative to the largest cluster within the dataset. Samples shown in the last row come from a cluster of size 2344, and is the largest among all clusters.

D. Waymo Open dataset

(a) $\mathcal{L}_{s(i)} = 0.2487$

(b) $\mathcal{L}_{s(i)} = 0.3178$

(c) $\mathcal{L}_{s(i)} = 0.7123$

(d) $\mathcal{L}_{s(i)} = 0.7462$

(e) $\mathcal{L}_{s(i)} = 1.0$

Figure S4: Samples from various clusters in Waymo *training* dataset, and their likelihoods relative to the largest cluster within the dataset. Totally, 1408 clusters formed within the 200000 training samples from the front camera. The rarest clusters contain (a) (b) and (c) of only 10 samples each in the empty residential driving scene, after-sunset scene, and high rainy scene respectively. Samples shown in the last row come from cluster 87 of size 796, and is the largest among all clusters. The rest top largest clusters share similar cluster sizes of 500 to 700 samples from urban city driving scenes, which indicates the Waymo training set is very well balanced.

E. BDD100K training dataset

(a) $\mathcal{L}_{s(i)} = 0.0049$

(b) $\mathcal{L}_{s(i)} = 0.0083$

(c) $\mathcal{L}_{s(i)} = 0.0086$

(d) $\mathcal{L}_{s(i)} = 0.7268$

(e) $\mathcal{L}_{s(i)} = 1.0$

Figure S5: Samples from various clusters in BDD100K *training* dataset, and their likelihoods relative to the largest cluster within the dataset. Samples shown in the last row come from a cluster of size 23385, and is the largest among all clusters. The smallest cluster shown in the top row contains only 116 samples.