LIGHTLY

Dataset Filtering and Analytics Report

25/01/2021 15:53:25

General Information

Job Information

Metric	Value
Build Time	Fri Jan 22 15:23:24 UTC 2021
Sampling Method	Coreset Algorithm
Number of Images	7481
Number of Corrupt Images	0
Number of Duplicates	0
Number of Removed Images	392
Number of Output Images	7089
Job Submitted	25/01/2021 15:48:01
Job Finished	25/01/2021 15:49:40
Total Processing Time	01m 39s

Estimated Savings

Task	Annotation Savings*	CO2 Savings*
Image Classification	$ 994.50	0.22 kg
Object Detection	$ 3978.00	0.80 kg
Semantic Segmentation	$ 19890.00	14.25 kg

*https://lightly.ai/report

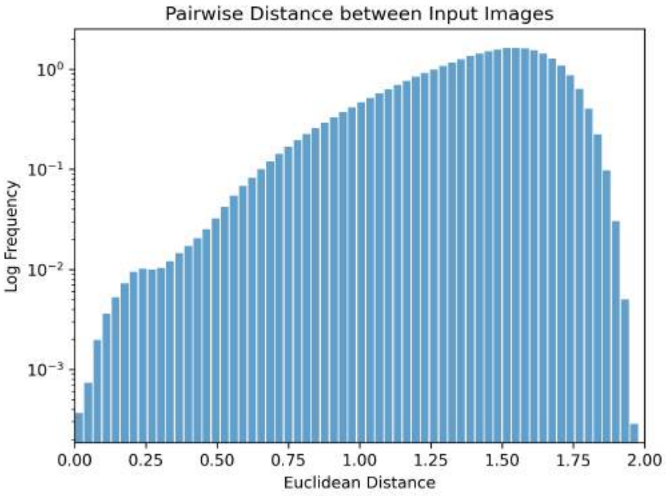
Statistics

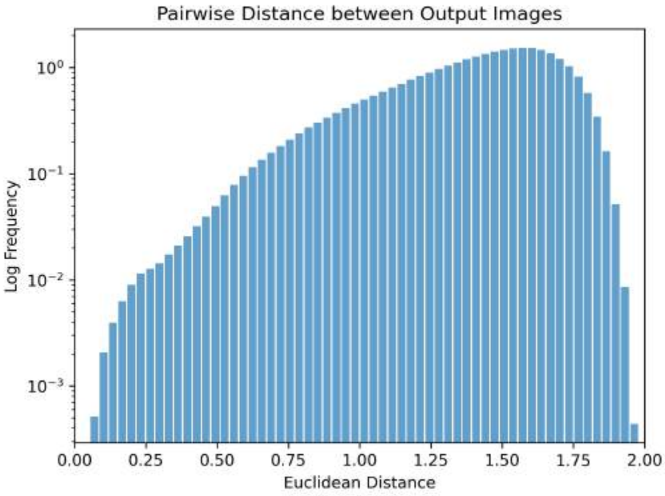
Metric	Before	After
Euclidean Distance (Mean)	1.3843	1.3811
Euclidean Distance (Min)	0.0000	0.0527
Euclidean Distance (Max)	1.9791	1.9803
Euclidean Distance (10th Percentile)	0.9952	0.9552
Euclidean Distance (90th Percentile)	1.6977	1.7169

Visualizations

Image Similarity in Input and Output Data

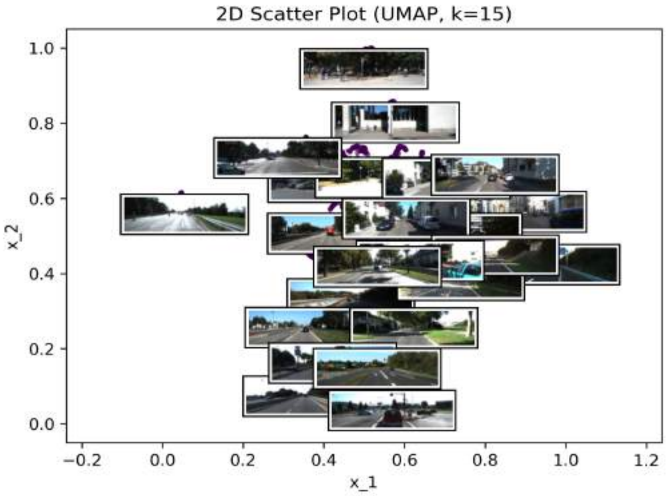
The plots below show the distribution of the pairwise distance between images in the input and output data. The histograms allow you to get information about the diversity of the dataset and whether the filter strength is well-chosen.

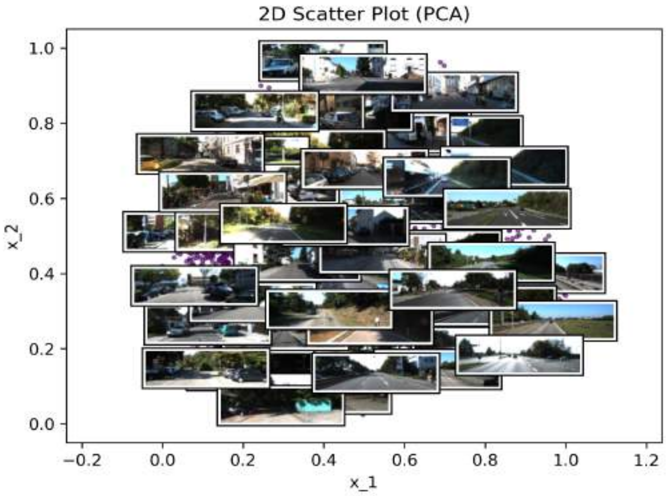




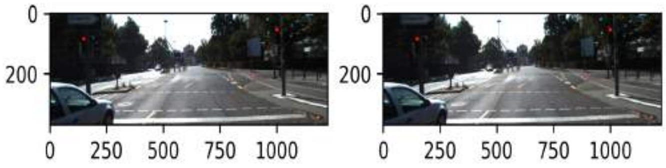
2D Scatter Plots of Output Data

Two-dimensional scatter plots help to understand the distribution of the data and may enable quick insights about outlier cases, dataset bias, or class imabalances.

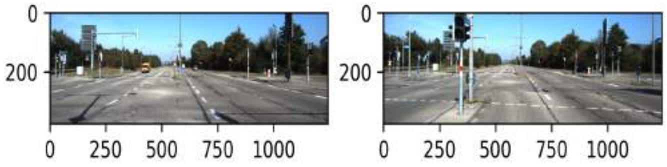




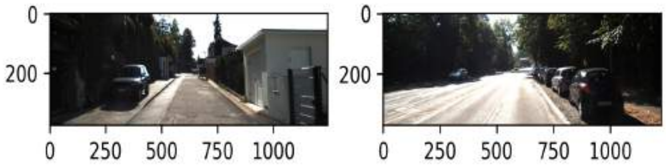
Retained (Left) and Removed (Right) Image with d = 0.01



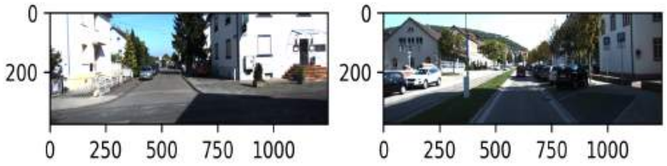
Retained (Left) and Removed (Right) Image with d = 0.27



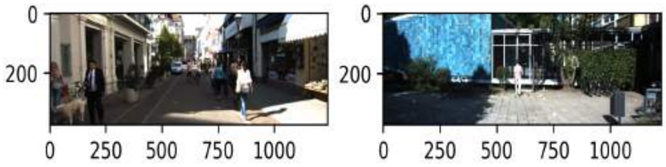
Retained (Left) and Removed (Right) Image with d = 0.45



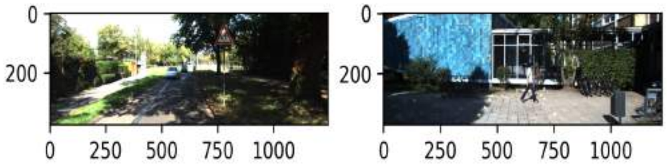
Retained (Left) and Removed (Right) Image with d = 0.55



Retained (Left) and Removed (Right) Image with d = 0.64



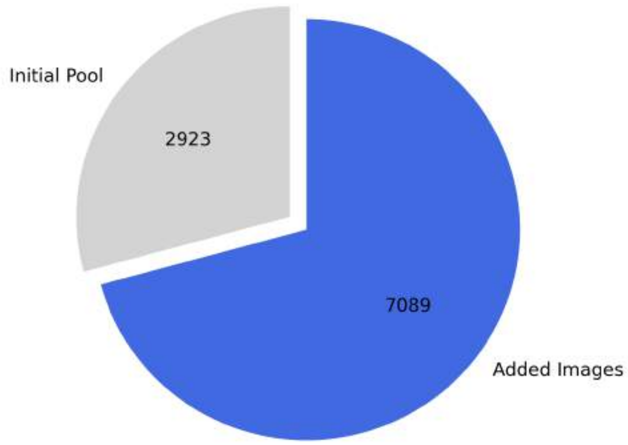
Retained (Left) and Removed (Right) Image with d = 0.75

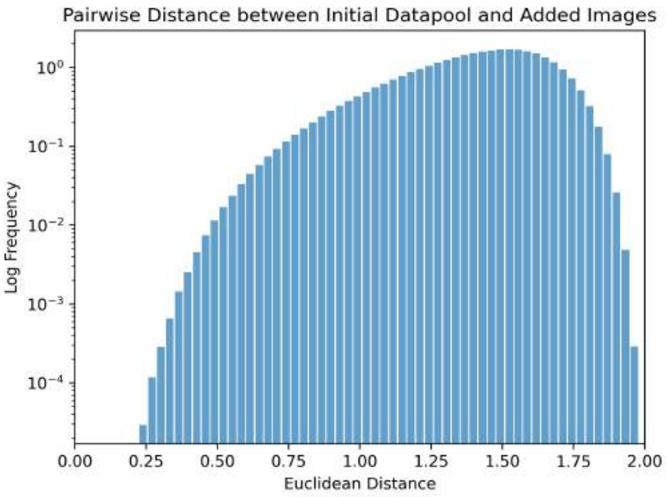


Datapool 1/2

Proportion of Added Images and Pairwise Distances

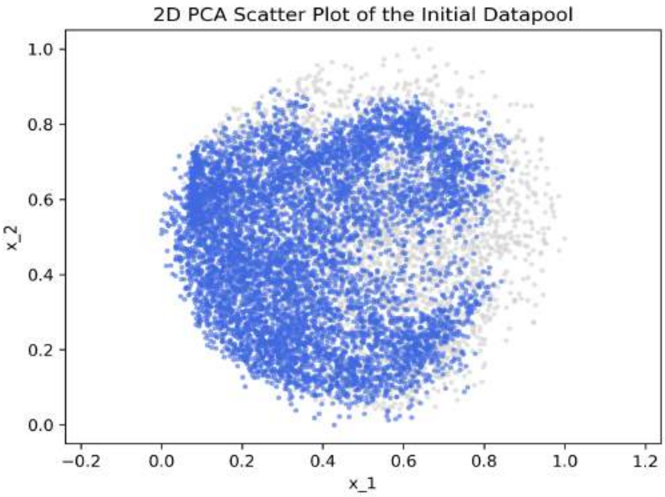
The figures below help to understand how many new images were added to the datapool and how similar the new images are to the ones selected in previous iterations.

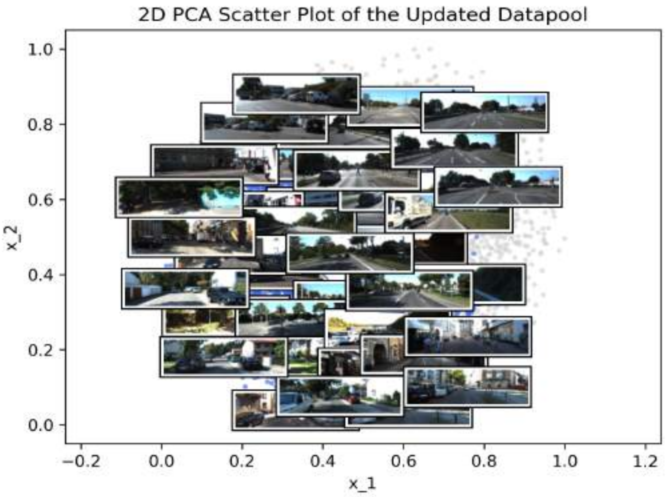




2D Scatter Plots of the Datapool (PCA)

The two-dimensional scatter plots of the datapool give an overview over the images which were added to it.

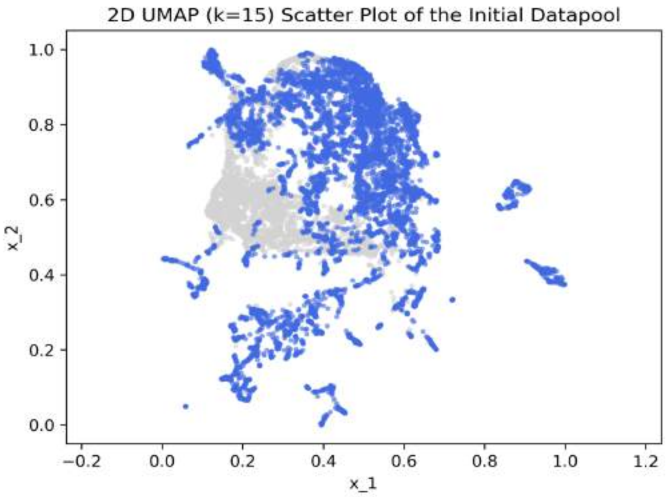




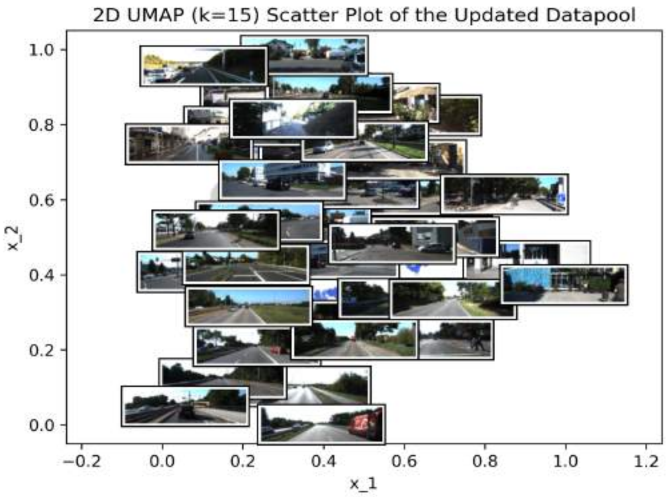
Datapool 2/2

2D Scatter Plots of the Datapool (UMAP, k=15)

The two-dimensional scatter plots of the datapool give an overview over the images which were added to it.



2D UMAP (k=15) Scatter Plot of the Initial Datapool



2D UMAP (k=15) Scatter Plot of the Updated Datapool