LIGHTLY

Dataset Filtering and Analytics Report

21/12/2023 09:42:57

LIGHTLY

General Information

Worker & Run Information

Metric	Value
Build Time	Wed Dec 20 14:27:15 UTC 2023
Build Version	2.10.dev
Job Submitted	2023-12-21 09:35:28
Job Finished	2023-12-21 09:42:20
Total Processing Time	06m 52s
Run ID	658406e096320fd94928ee08
Dataset ID	658406c01a25e8ee0795cc98
Dataset Name	test_volleyball_object_frequency_2023-12-21-09-34-54-976231

All this data and more is also available in the report_v2.json file.

Dataset Sizes

Metric	Image	Video
Input	3756	5
Corrupt	0	0
Duplicate	0	0
Removed	3656	5
Selected	100	5
Datapool Input	0	0
Datapool Selected	100	5

A video is considered {corrupt, ...} if it contains any {corrupt, ...} frames.

More information about corrupt frames and videos is also found in the corruptness_check_information.json file.

General Information

Estimated Savings

Task	Cost Savings	CO2 Savings
Image Classification	$ 1096.80	0.11 kg
Object Detection	$ 4387.20	0.40 kg
Semantic Segmentation	$ 21936.00	7.68 kg

* https://lightly.ai/report

Selection Results

This page shows statistics of the selected data. All results are compared between the selection made by Lightly and a random selection. The results are calculated including the datapool if this dataset has a datapool. More details on the different metrics can be found in our docs: https://docs.lightly.ai/docs/dataset-metrics.

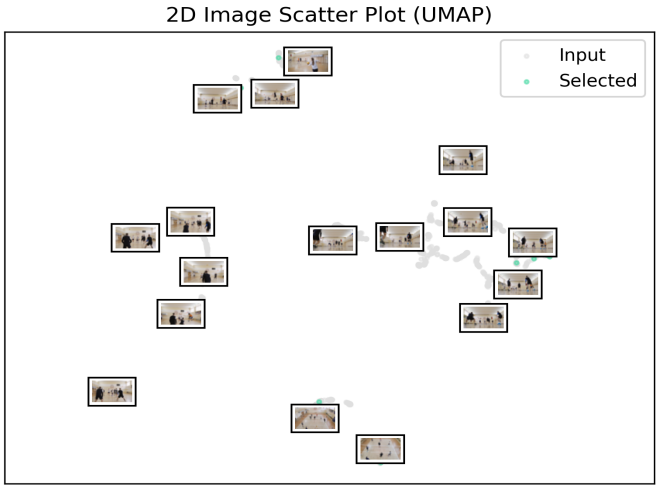
Random and Lightly Selection Results

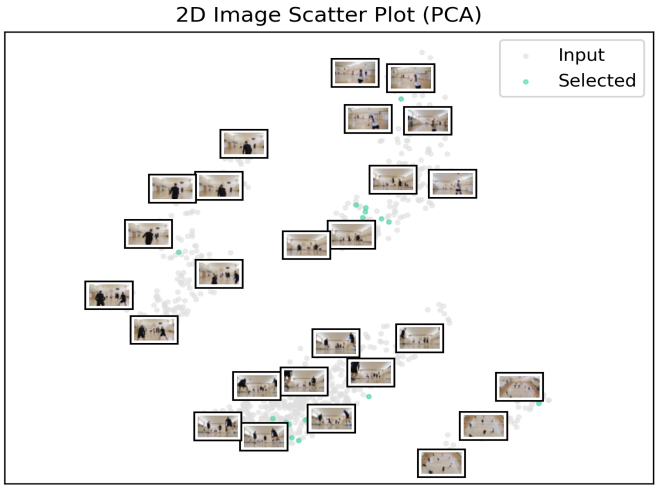
Metric	Random	Lightly	Improvement Over Random
Image Diversity	0.109	0.054	-50.7%
Image Coverage	0.916	0.872	-4.8%
Object Diversity	1.243	0.895	-28%
Object Coverage	0.459	0.439	-4.3%
Object Balance	1	1	0%

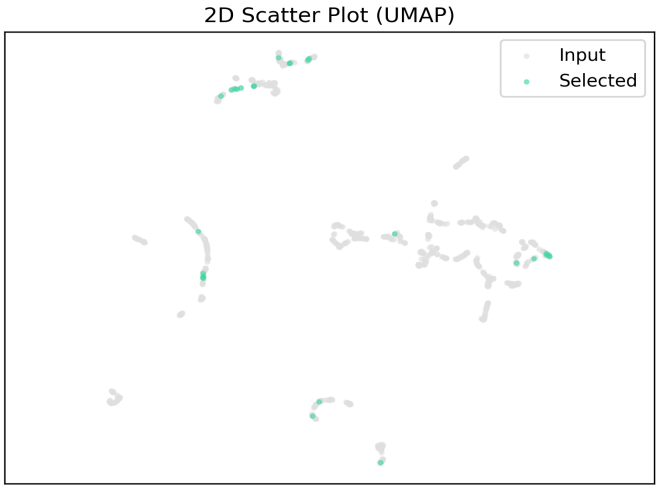
Image Level Analysis - Embeddings

Embedding 2D Scatter Plots

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







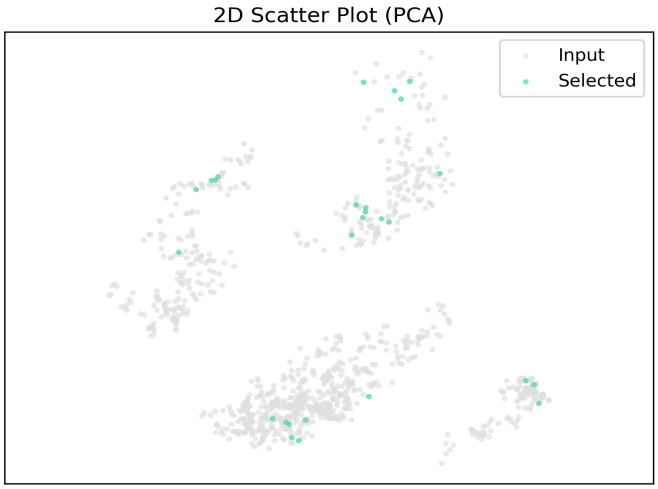
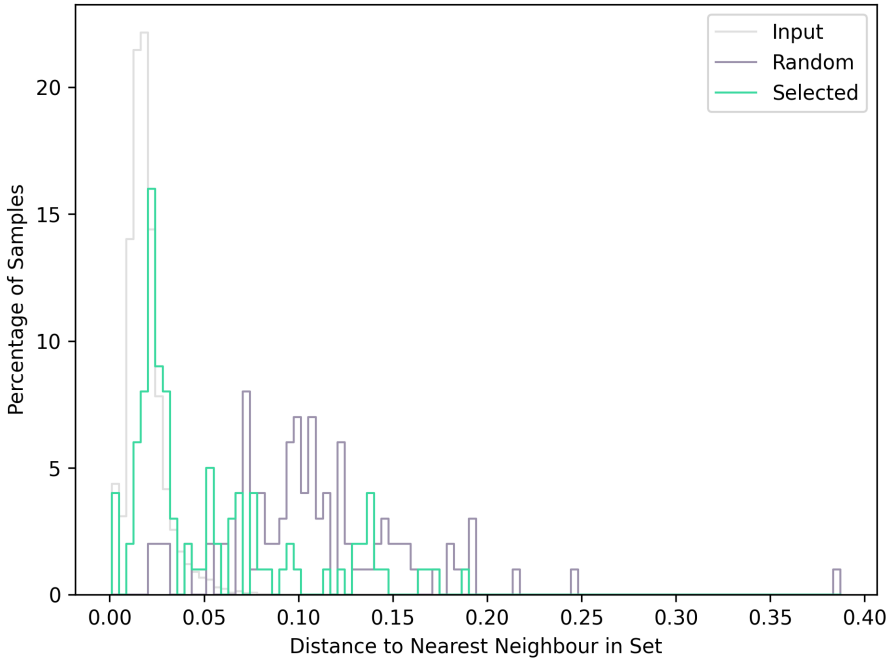


Image Level Analysis - Embeddings

Embedding Diversity Metric

For the diversity metrics, we compute the distance from each sample to its closest neighbour sample in the same set. Higher diversity means lower information redundancy in the dataset. For a detailed explanation of the metric, see our docs. To improve this metric, use diversity selection.



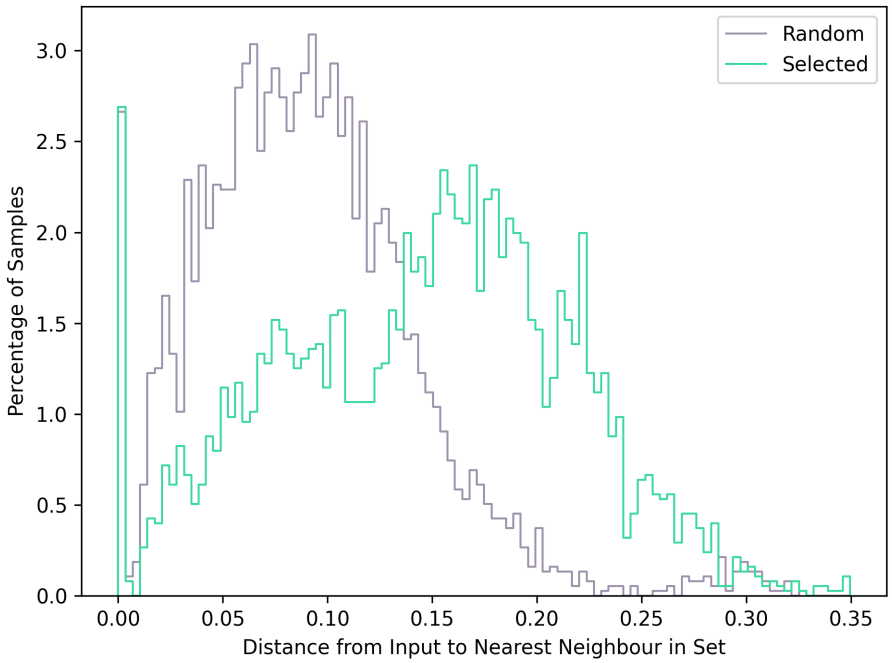
Distances to Nearest Neighbour Within Same Set

Set	Mean	Std	Min	Median	Max
Input	0.019	0.010	0.001	0.018	0.085
Random	0.109	0.051	0.022	0.104	0.387
Selected	0.054	0.046	0.001	0.030	0.187

Image Level Analysis - Embeddings

Embedding Coverage Distance Metric

The coverage measures how well the input set is covered by a subset of it. It is computed as the distance from each input sample to the closest sample in the subset. Low values mean the selected samples cover the input space well as for each not selected sample, there's at least one selected sample that is similar. For a detailed explanation of the metric, see our docs.

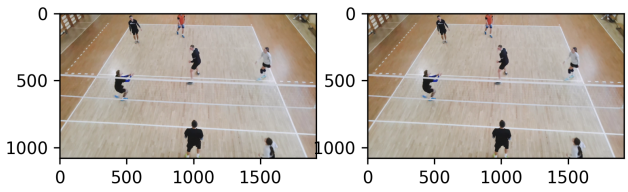


Distances from Input to Nearest Neighbour in Set

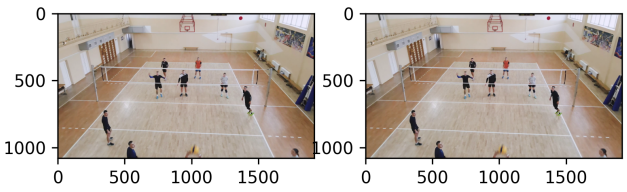
Set	Mean	Std	Min	Median	Max
Random	0.091	0.052	0	0.087	0.325
Selected	0.146	0.070	0	0.153	0.350

Sample of Selected Images and Similar Removed Images

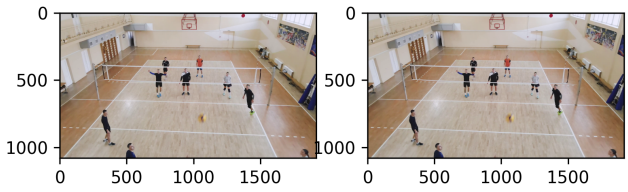
Retained (Left) and Removed (Right) Image with d = 0.00



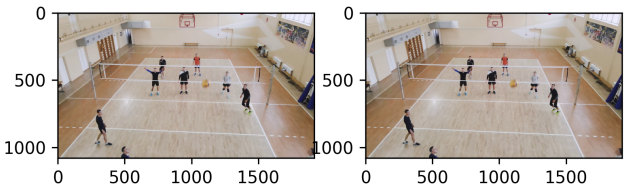
Retained (Left) and Removed (Right) Image with d = 0.00



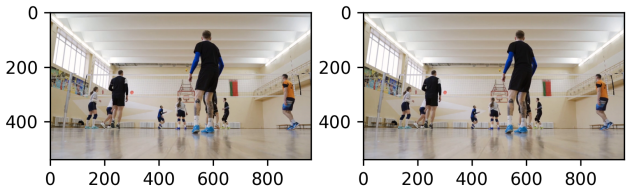
Retained (Left) and Removed (Right) Image with d = 0.01



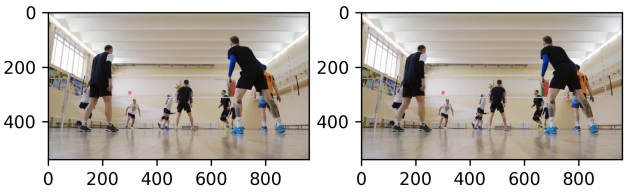
Retained (Left) and Removed (Right) Image with d = 0.01



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

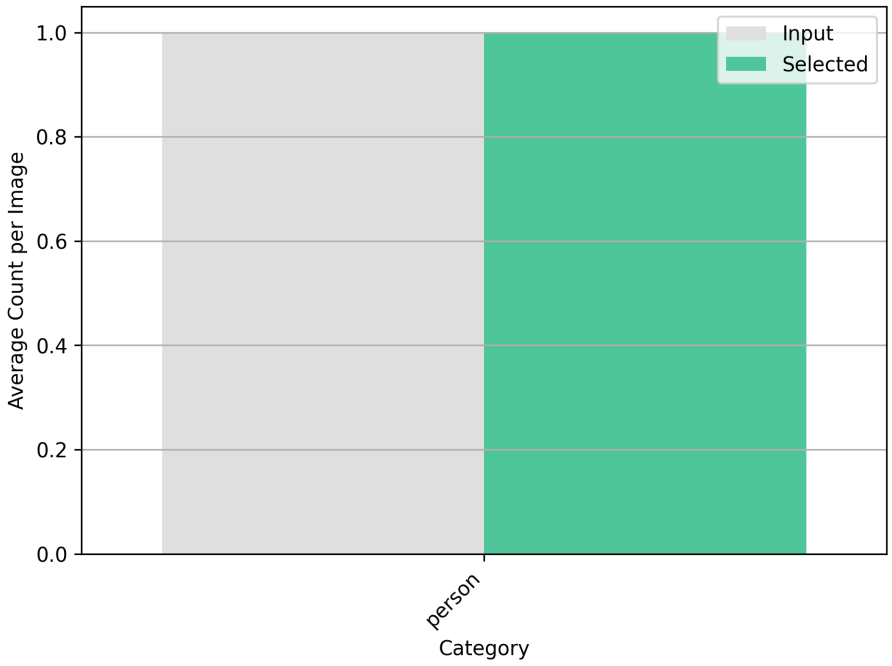


Retained (Left) and Removed (Right) Image with d = 0.01



Prediction Task: keypoint_rcnn_detection

Average Category Counts per Image



Average Category Counts per Image

Category	Input	Selected
person	16.007	25.850
All Categories	16.007	25.850

Prediction Task: keypoint_rcnn_detection

Category Distribution

Category	Input	Selected
person	100%	100%
All Categories	100%	100%

Total Category Counts

Category	Input	Selected
person	60122	2585
All Categories	60122	2585

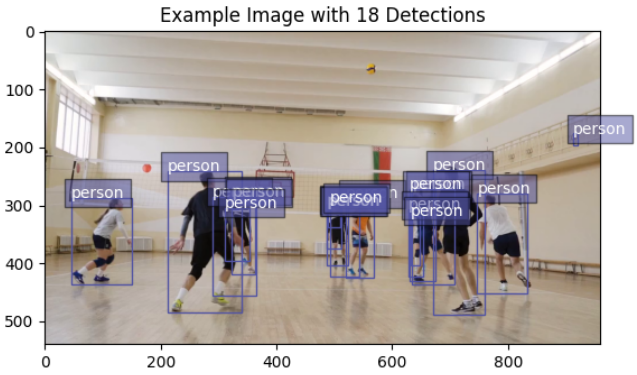
Prediction Task: keypoint_rcnn_detection

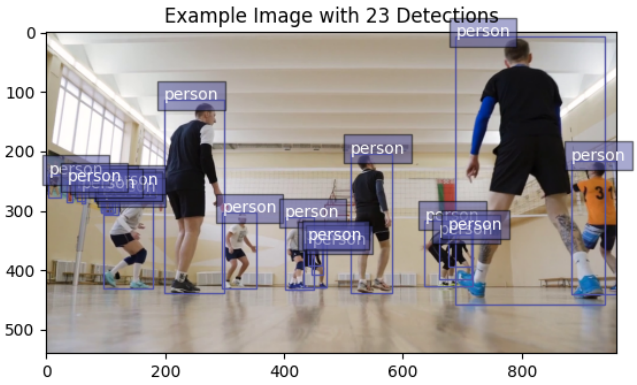
Categorical Metrics

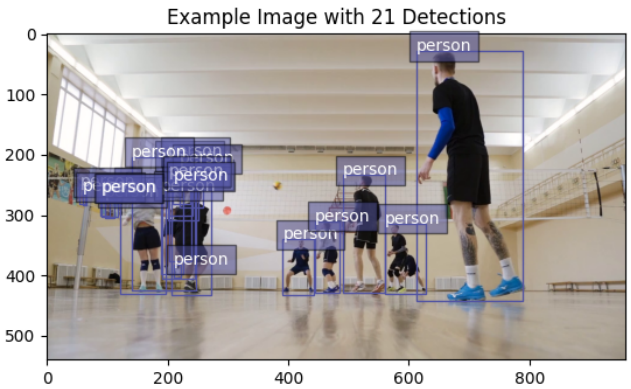
Metric	Input	Selected
Dataset Balance	1	1

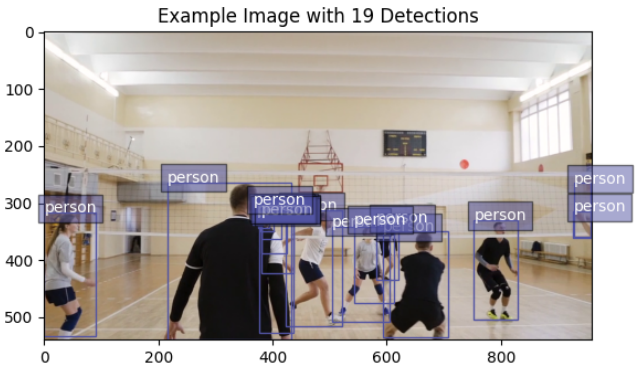
We measure the dataset balance via the normalized entropy. The higher the score, the more balanced the dataset. It is 0 if one category has all counts and 1 if all categories have the same count.

Prediction Task: keypoint_rcnn_detection





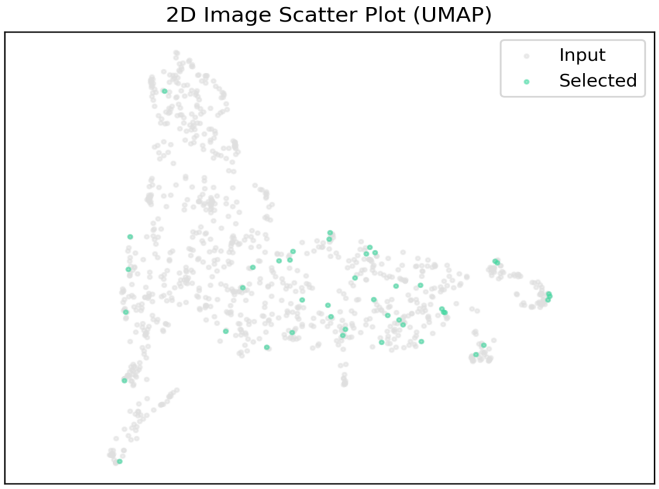


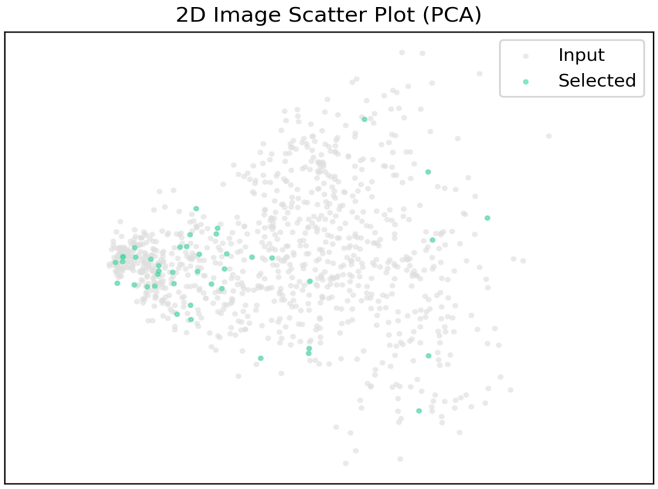


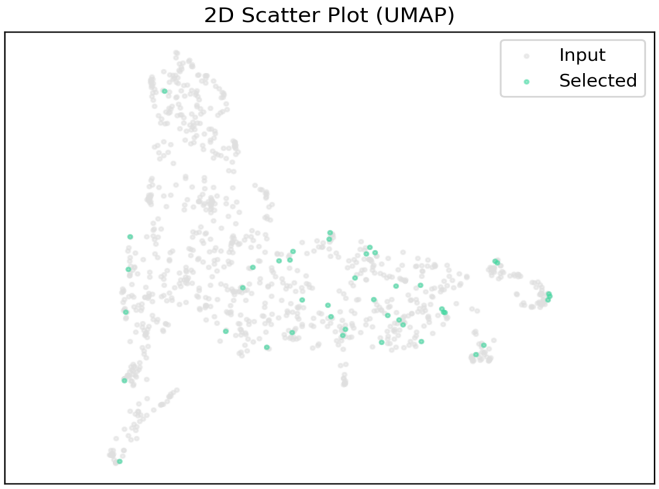
Prediction task: keypoint_rcnn_detection

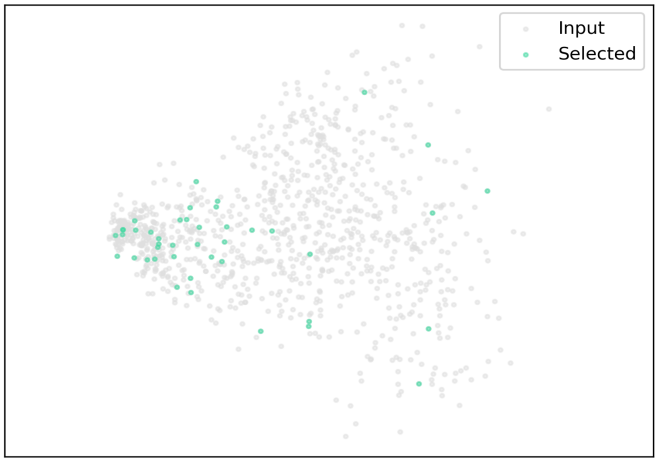
Crop Embedding 2D Scatter Plots

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





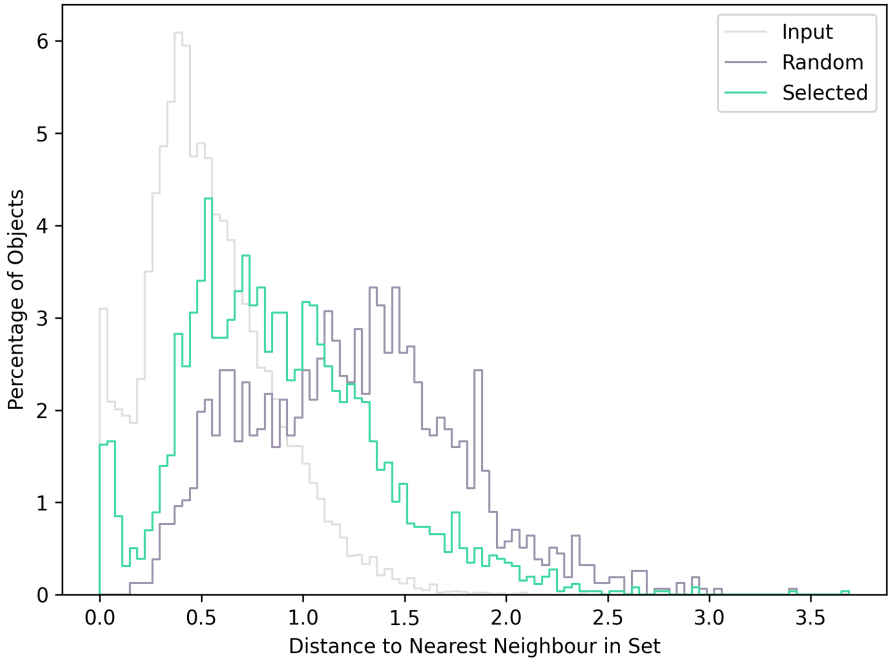




Prediction task: keypoint_rcnn_detection

Embedding Diversity Metric

For the diversity metrics, we compute the distance from each object to its closest neighbour object in the same set. Higher diversity means lower information redundancy in the dataset. For a detailed explanation of the metric, see our docs. To improve this metric, use diversity selection.



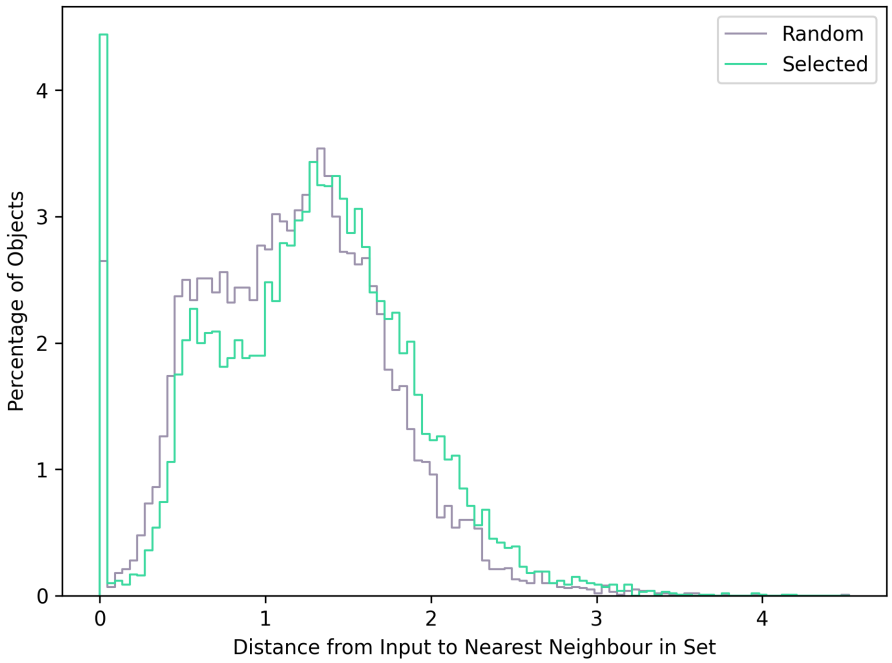
Distances to Nearest Neighbour Within Same Set

Set	Mean	Std	Min	Median	Max
Input	0.539	0.314	0	0.492	2.601
Random	1.243	0.520	0.156	1.241	3.418
Selected	0.895	0.474	0	0.840	3.688

Prediction task: keypoint_rcnn_detection

Embedding Coverage Distance Metric

The coverage measures how well the input set is covered by a subset of it. It is computed as the distance from each input object to the closest object in the subset. Low values mean the selected objects cover the input space well as for each not selected object, there's at least one selected object that is similar. For a detailed explanation of the metric, see our docs.

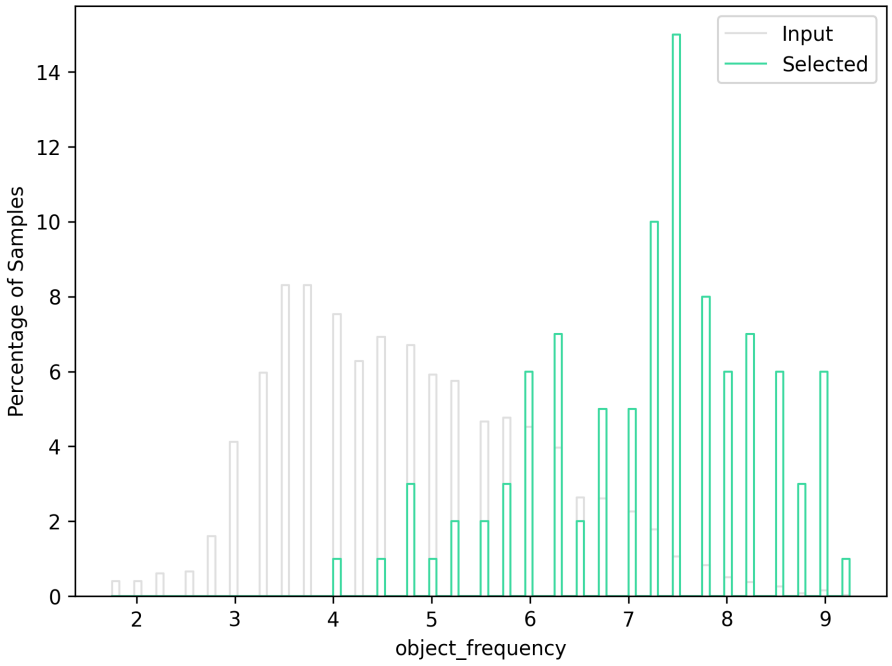


Distances from Input to Nearest Neighbour in Set

Set	Mean	Std	Min	Median	Max
Random	1.179	0.563	0	1.183	4.525
Selected	1.276	0.612	0	1.300	4.207

Prediction task: keypoint_rcnn_detection

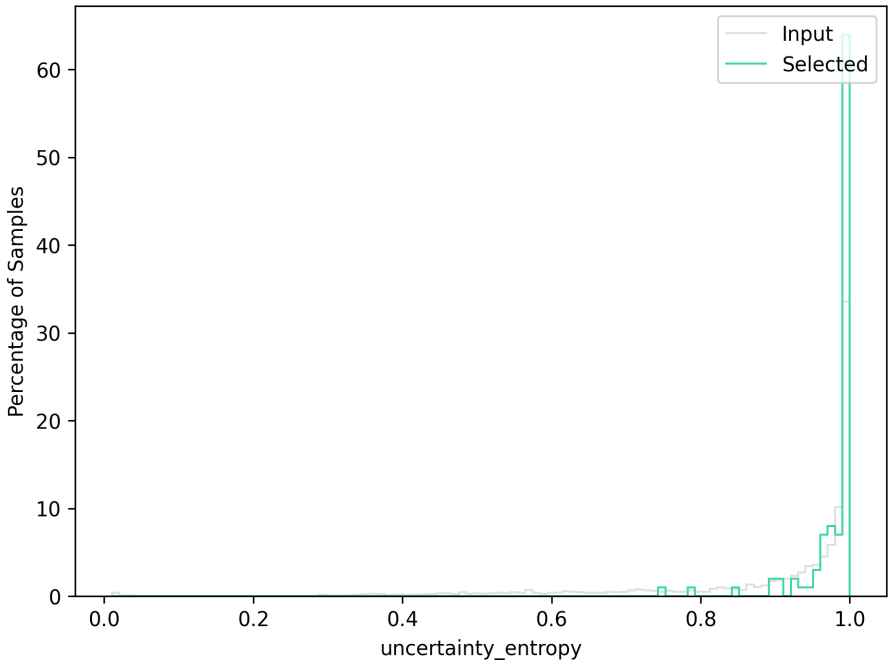
Active Learning Score: object_frequency



Set	Mean	Std	Min	Median	Max
Input	4.752	1.319	1.750	4.500	9.250
Selected	7.213	1.156	4	7.500	9.250

Prediction task: keypoint_rcnn_detection

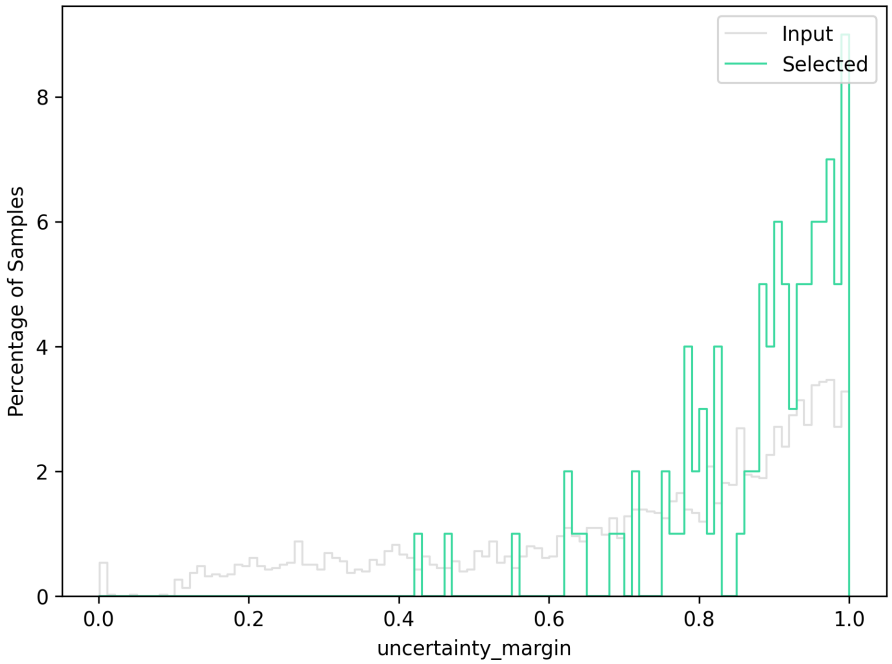
Active Learning Score: uncertainty_entropy



Set	Mean	Std	Min	Median	Max
Input	0.889	0.171	0.010	0.969	1
Selected	0.979	0.041	0.746	0.995	1

Prediction task: keypoint_rcnn_detection

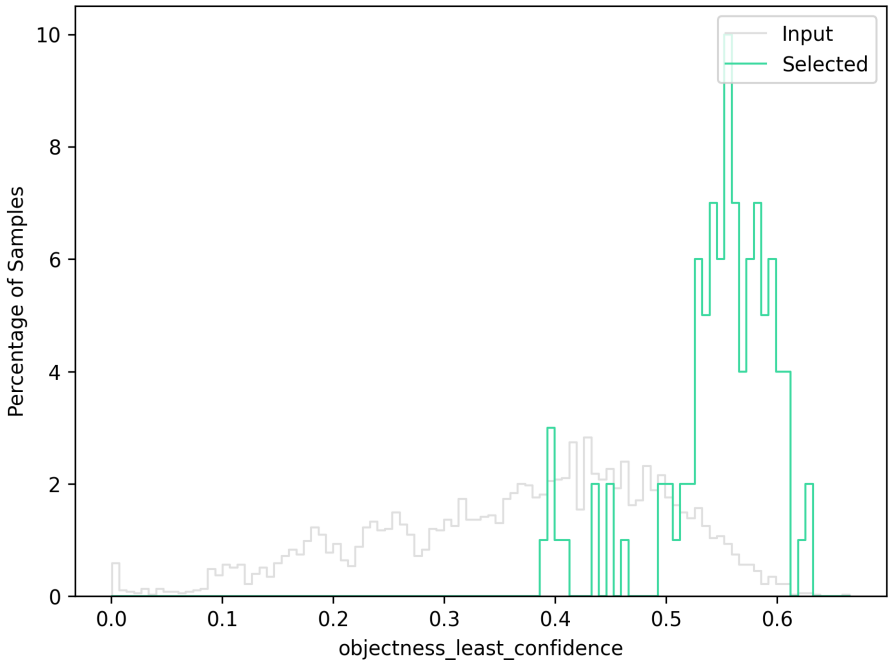
Active Learning Score: uncertainty_margin



Set	Mean	Std	Min	Median	Max
Input	0.718	0.246	0.002	0.795	1
Selected	0.879	0.117	0.425	0.916	0.998

Prediction task: keypoint_rcnn_detection

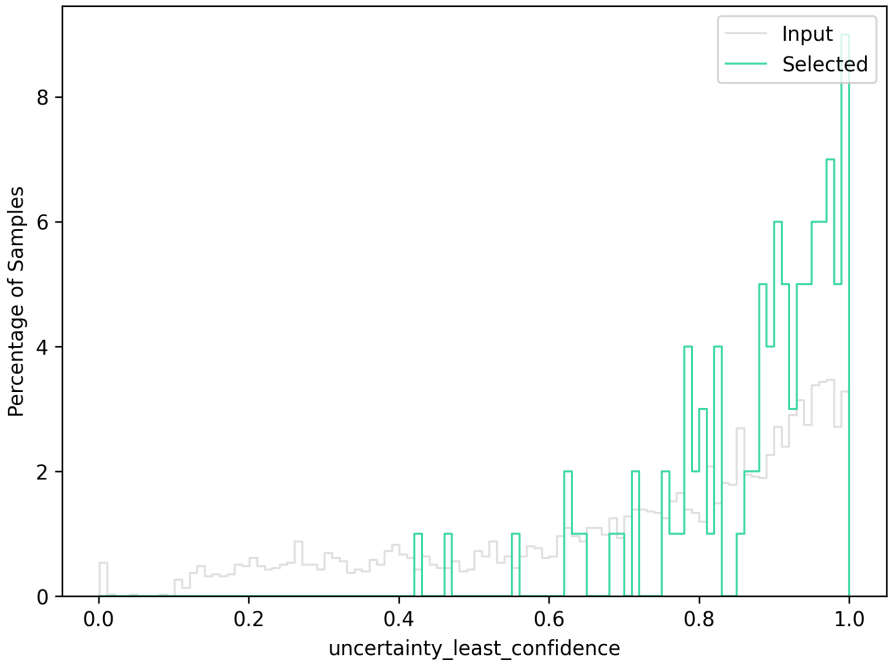
Active Learning Score: objectness_least_confidence



Set	Mean	Std	Min	Median	Max
Input	0.370	0.128	0.001	0.390	0.665
Selected	0.547	0.054	0.391	0.556	0.629

Prediction task: keypoint_rcnn_detection

Active Learning Score: uncertainty_least_confidence

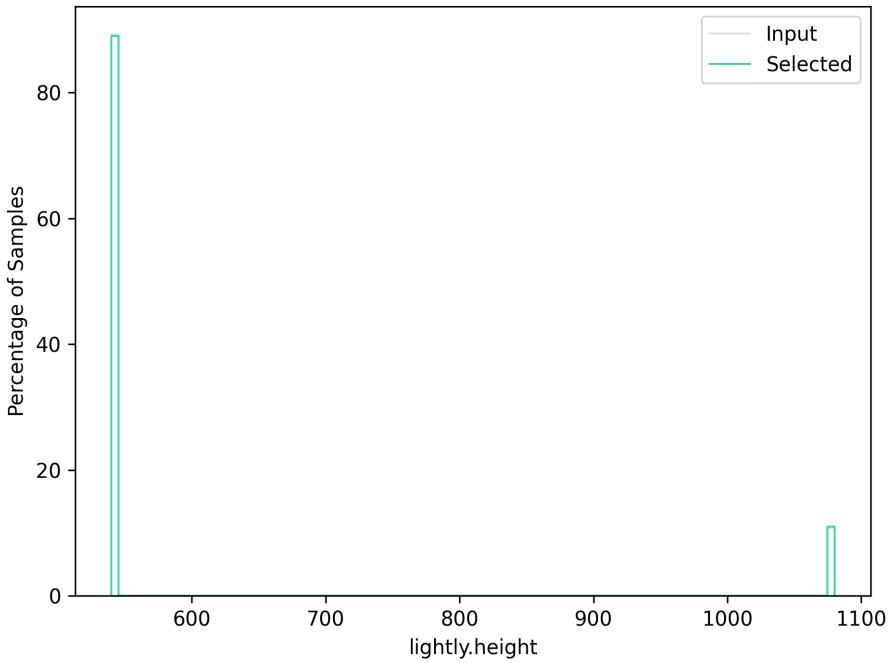


Set	Mean	Std	Min	Median	Max
Input	0.718	0.246	0.002	0.795	1
Selected	0.879	0.117	0.425	0.916	0.998

Lightly Metadata

Lightly provides a set of metadata for each image. This metadata can be used in the selection process using the THRESHOLD or WEIGHTS strategy. Read more here: https://docs.lightly.ai/docs/selection#metadata

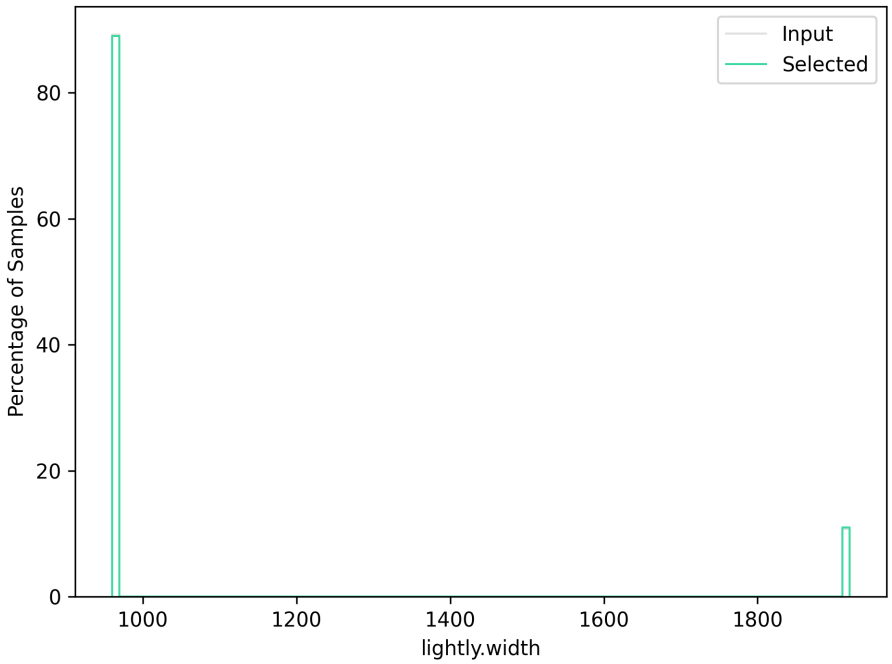
Input Height (lightly.height)



Set	Mean	Std	Min	Median	Max
Input	598.227	167.488	540	540	1080
Selected	599.400	168.960	540	540	1080

Lightly Metadata

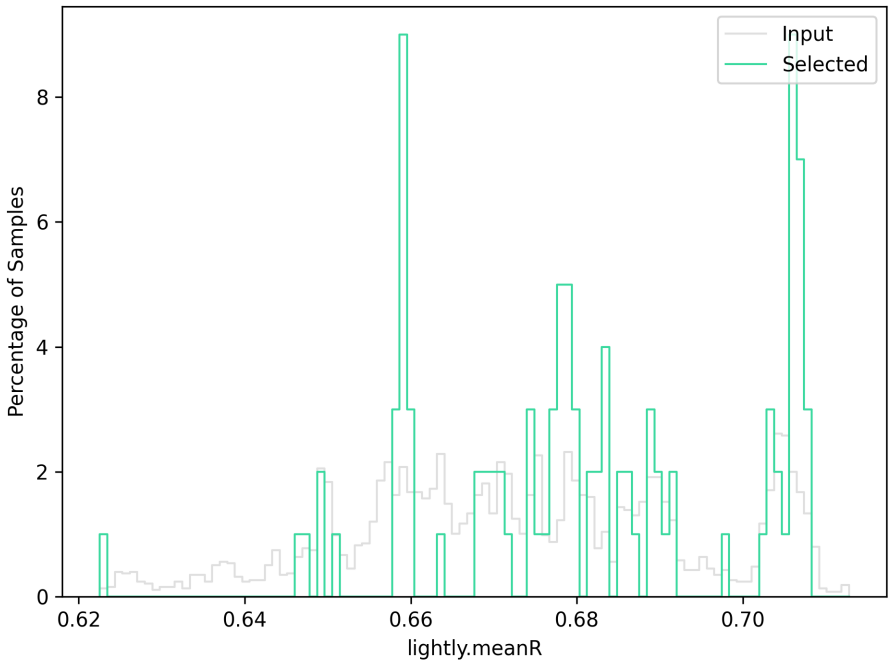
Input Width (lightly.width)



Set	Mean	Std	Min	Median	Max
Input	1063.514	297.756	960	960	1920
Selected	1065.600	300.374	960	960	1920

Lightly Metadata

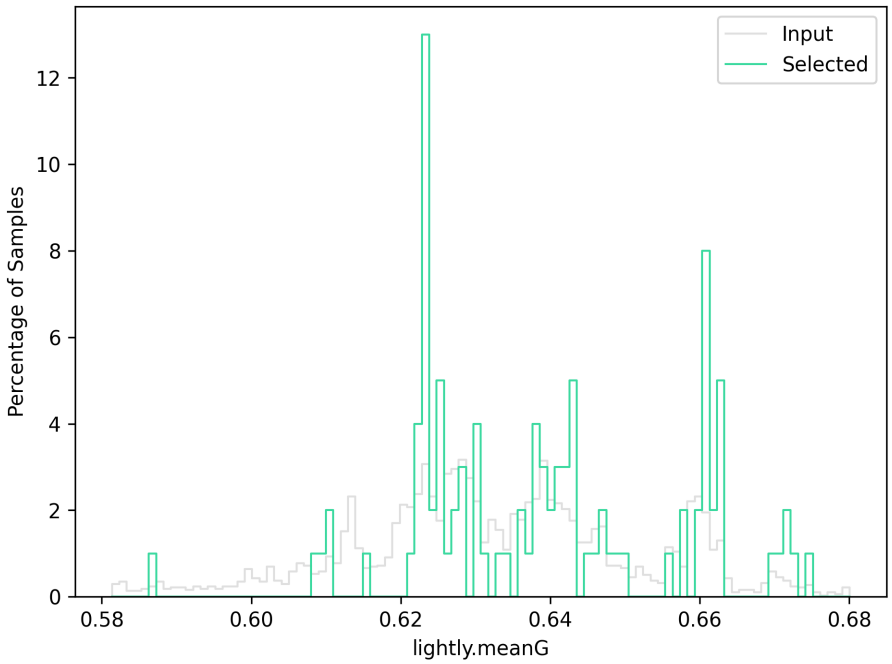
Red Channel Mean (lightly.meanR)



Set	Mean	Std	Min	Median	Max
Input	0.673	0.021	0.623	0.672	0.713
Selected	0.681	0.019	0.623	0.680	0.708

Lightly Metadata

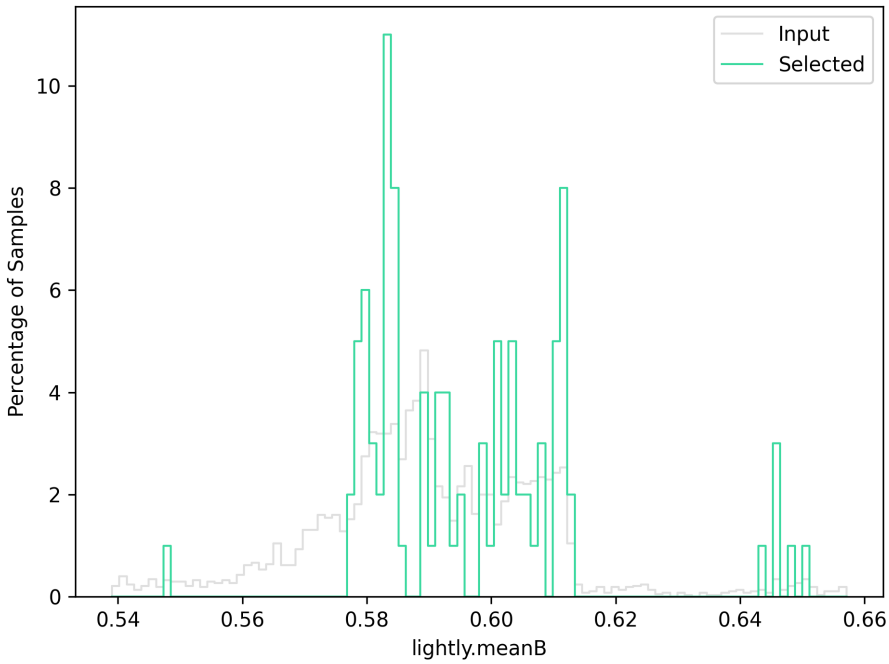
Green Channel Mean (lightly.meanG)



Set	Mean	Std	Min	Median	Max
Input	0.633	0.019	0.581	0.632	0.680
Selected	0.639	0.018	0.586	0.638	0.675

Lightly Metadata

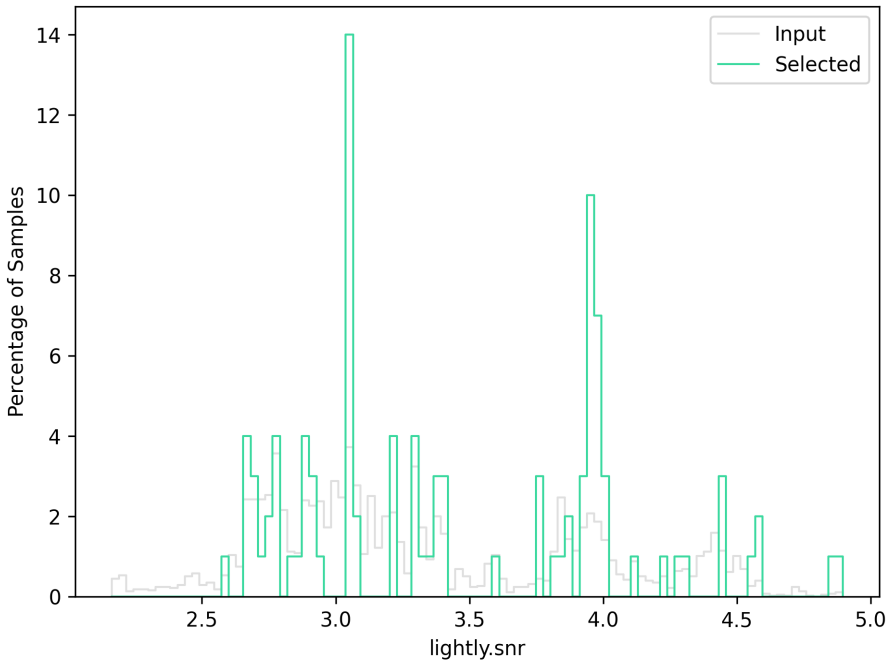
Blue Channel Mean (lightly.meanB)



Set	Mean	Std	Min	Median	Max
Input	0.590	0.018	0.539	0.589	0.657
Selected	0.597	0.018	0.548	0.593	0.650

Lightly Metadata

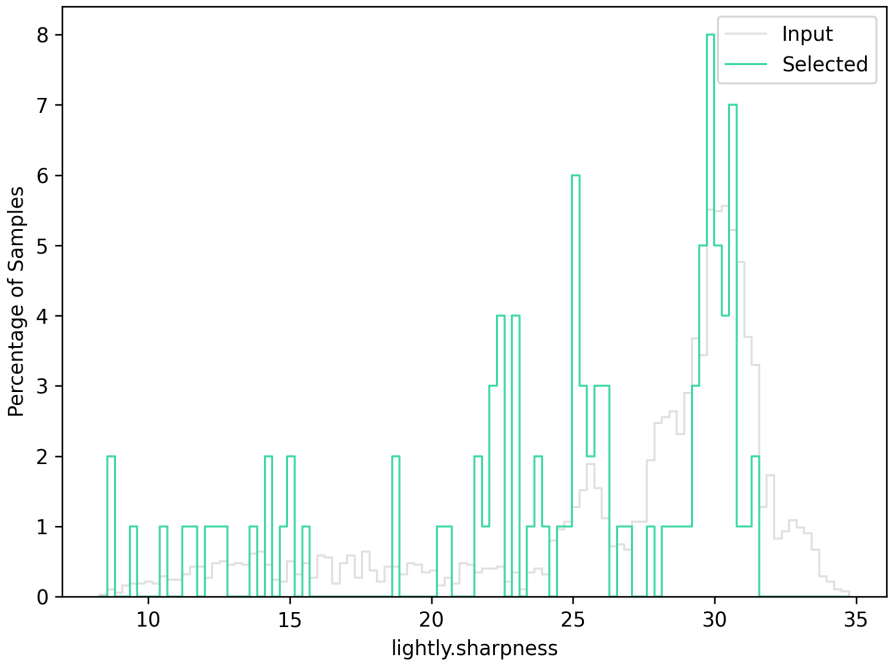
Signal to Noise Ratio (lightly.snr)



Set	Mean	Std	Min	Median	Max
Input	3.331	0.596	2.163	3.174	4.895
Selected	3.468	0.586	2.598	3.330	4.888

Lightly Metadata

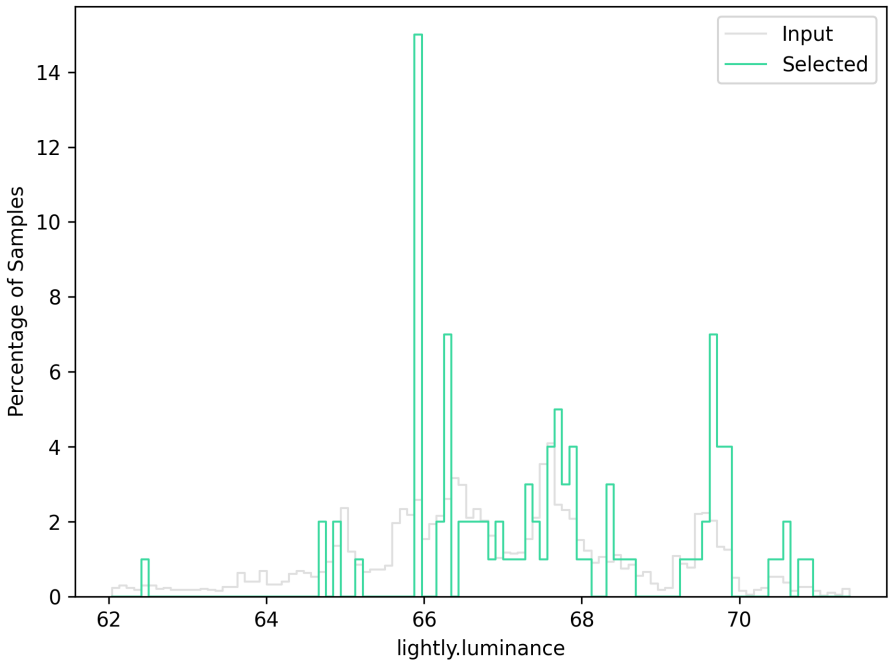
Sharpness (lightly.sharpness)



Set	Mean	Std	Min	Median	Max
Input	26.756	5.849	8.292	29.183	34.750
Selected	24.527	6.199	8.595	25.655	31.474

Lightly Metadata

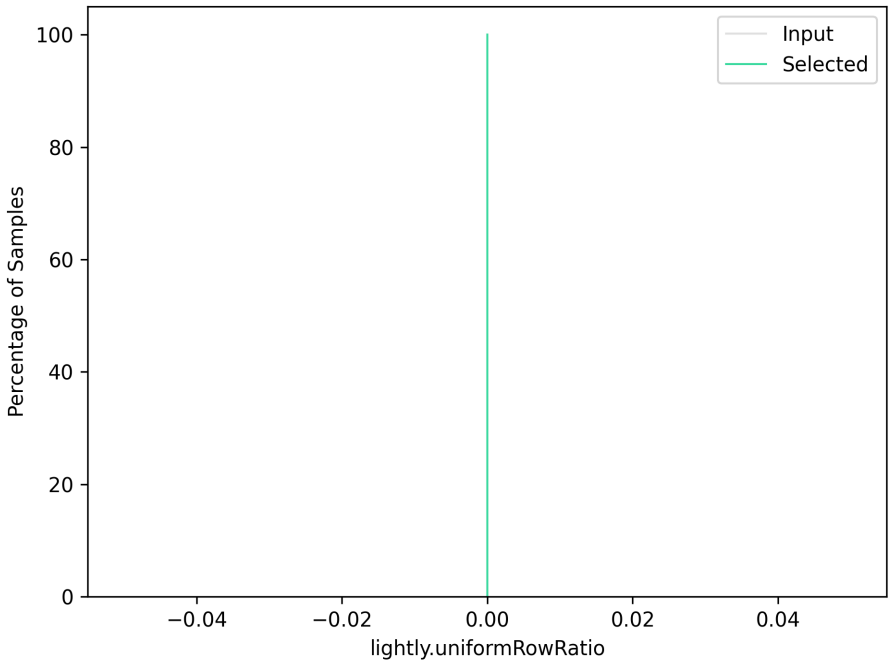
Luminance (lightly.luminance)



Set	Mean	Std	Min	Median	Max
Input	66.944	1.811	62.042	66.822	71.400
Selected	67.597	1.665	62.431	67.560	70.893

Lightly Metadata

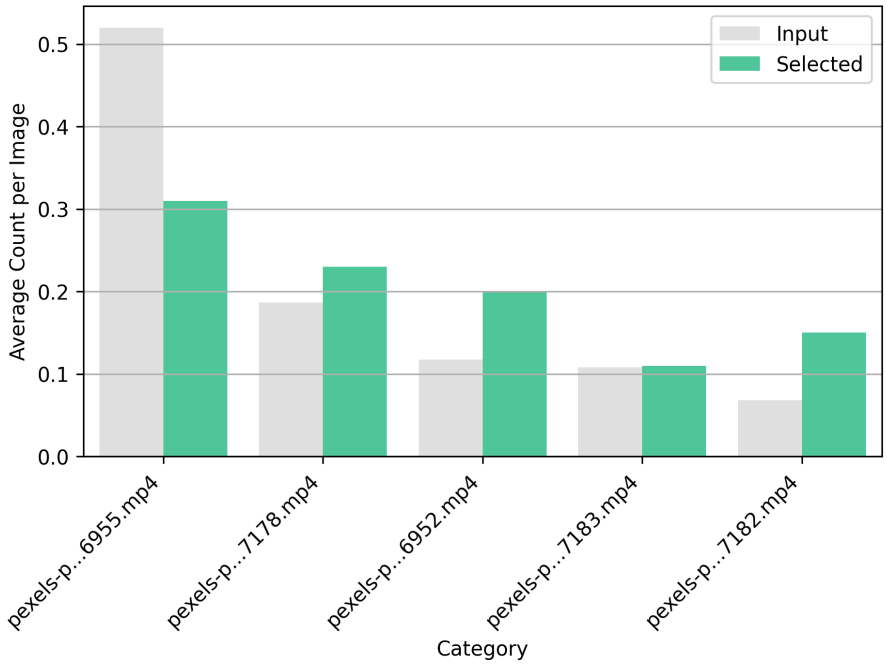
Uniform Row Ratio (lightly.uniformRowRatio)



Set	Mean	Std	Min	Median	Max
Input	0	0	0	0	0
Selected	0	0	0	0	0

Categorical Metadata: video_name

Video Name (video_name)



Average Category Counts per Image

Category	Input	Selected
pexels-p...6952.mp4	0.117	0.200
pexels-p...6955.mp4	0.520	0.310
pexels-p...7178.mp4	0.186	0.230
pexels-p...7182.mp4	0.068	0.150
pexels-p...7183.mp4	0.108	0.110
All Categories	1.000	1.000

Categorical Metadata: video_name

Category Distribution

Category	Input	Selected
pexels-p...6952.mp4	11.7%	20%
pexels-p...6955.mp4	52%	31%
pexels-p...7178.mp4	18.6%	23%
pexels-p...7182.mp4	6.8%	15%
pexels-p...7183.mp4	10.8%	11%
All Categories	100%	100%

Total Category Counts

Category	Input	Selected
pexels-p...6952.mp4	441	20
pexels-p...6955.mp4	1953	31
pexels-p...7178.mp4	700	23
pexels-p...7182.mp4	257	15
pexels-p...7183.mp4	405	11
All Categories	3756	100

Video Sampling Densities 1/1

We show selected frames for each video. Each selected frame is indicated by a vertical line. When using coreset, clusters of selected frames show sequences where the frames differ a lot visually. Additionally, high density regions will appear darker in the plots.