LIGHTLY

Dataset Filtering and Analytics Report

21/12/2023 08:26:11

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 08:21:25
Job Finished	2023-12-21 08:25:08
Total Processing Time	03m 43s
Run ID	6583f5859f4d5eca1dbe8e46
Dataset ID	6583f5789f4d5eca1dbe8e21
Dataset Name	test_diversity_bdd_video_2023-12-21-08-21-07-908454

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

Dataset Sizes

Metric	Image	Video
Input	7223	5
Corrupt	0	0
Duplicate	0	0
Removed	6501	5
Selected	722	5
Datapool Input	0	0
Datapool Selected	722	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	$ 1950.30	0.20 kg
Object Detection	$ 7801.20	0.72 kg
Semantic Segmentation	$ 39006.00	13.65 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.

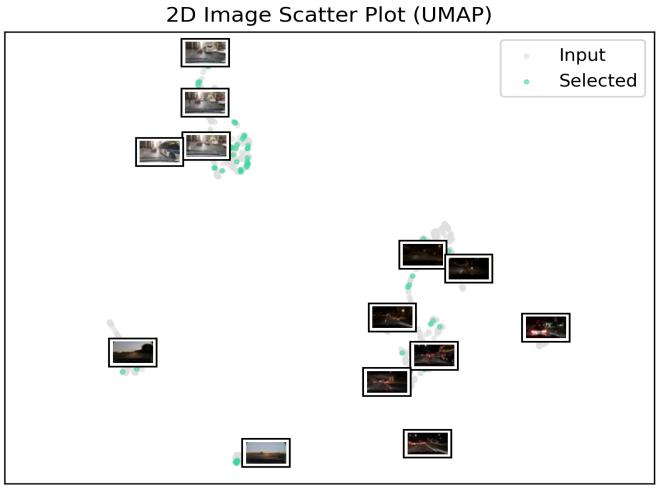
Random and Lightly Selection Results

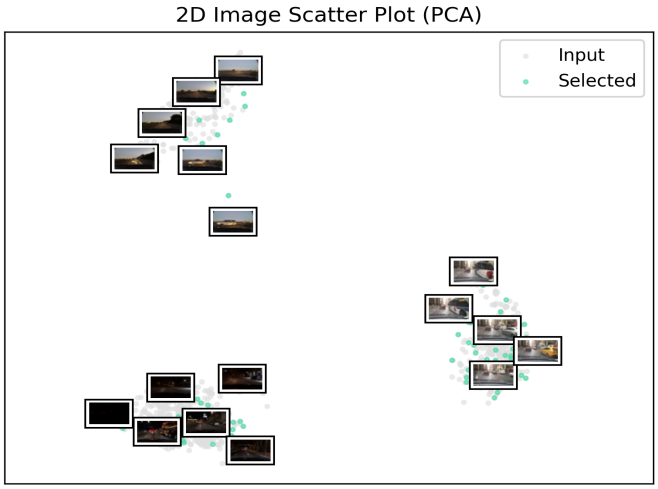
Metric	Random	Lightly	Improvement Over Random
Image Diversity	0.075	0.142	+88.9%
Image Coverage	0.936	0.937	+0.1%

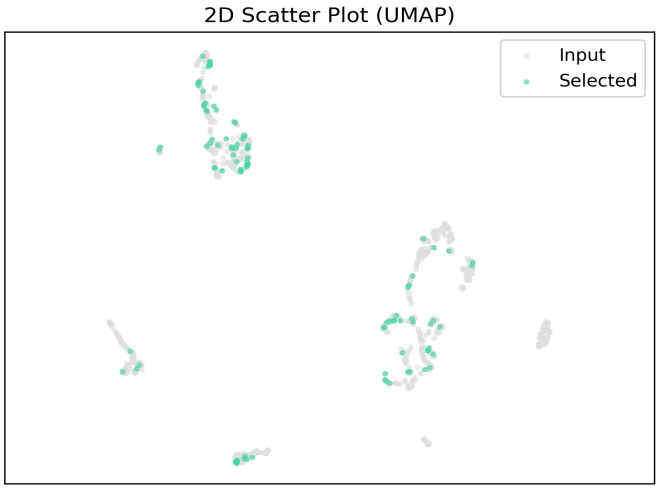
Image Level Analysis - Embeddings

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.







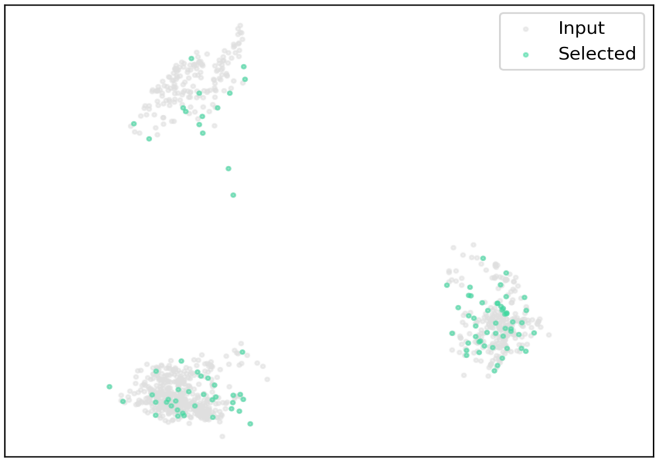
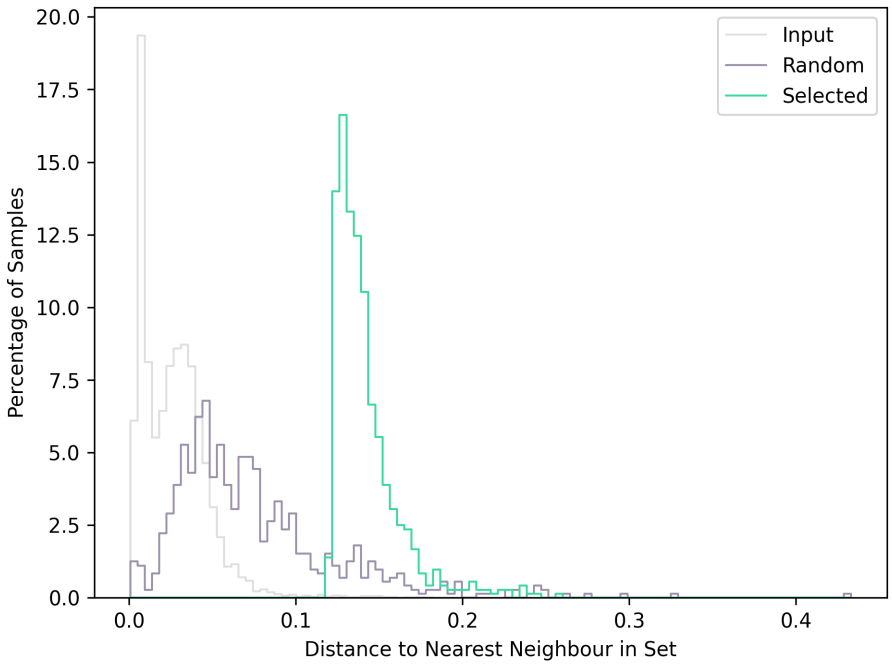


Image Level Analysis - Embeddings

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.



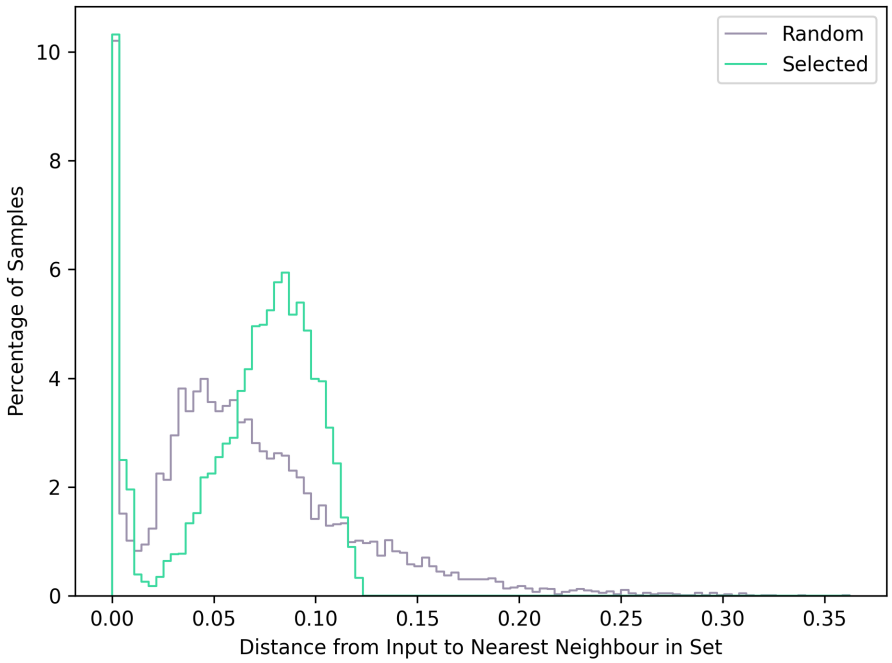
Distances to Nearest Neighbour Within Same Set

Set	Mean	Std	Min	Median	Max
Input	0.026	0.019	0.001	0.025	0.179
Random	0.075	0.050	0.001	0.064	0.433
Selected	0.142	0.020	0.122	0.137	0.257

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.

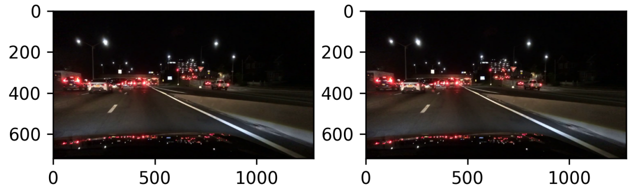


Distances from Input to Nearest Neighbour in Set

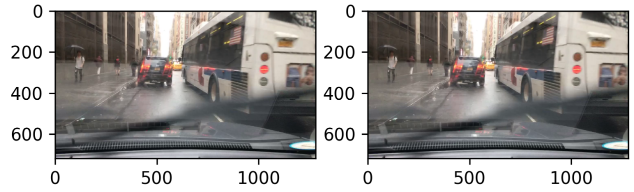
Set	Mean	Std	Min	Median	Max
Random	0.068	0.050	0	0.060	0.362
Selected	0.067	0.033	0	0.075	0.121

Sample of Selected Images and Similar Removed Images

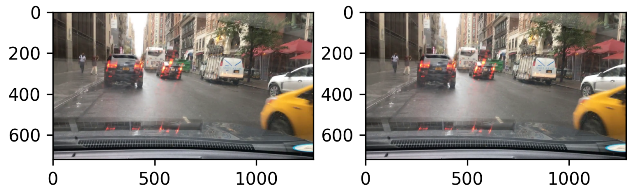
Retained (Left) and Removed (Right) Image with d = 0.00



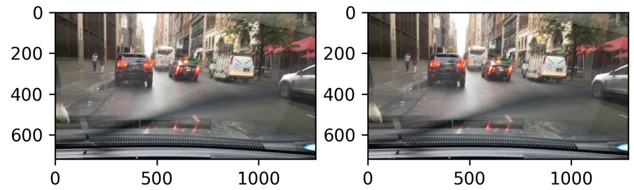
Retained (Left) and Removed (Right) Image with d = 0.00



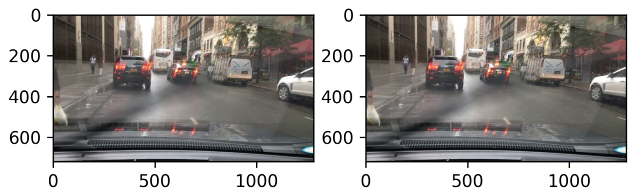
Retained (Left) and Removed (Right) Image with d = 0.00



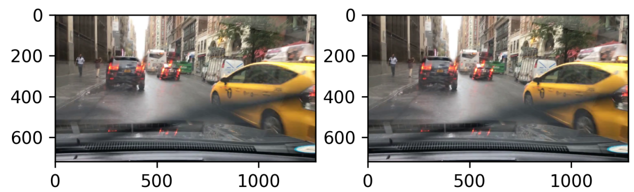
Retained (Left) and Removed (Right) Image with d = 0.00



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



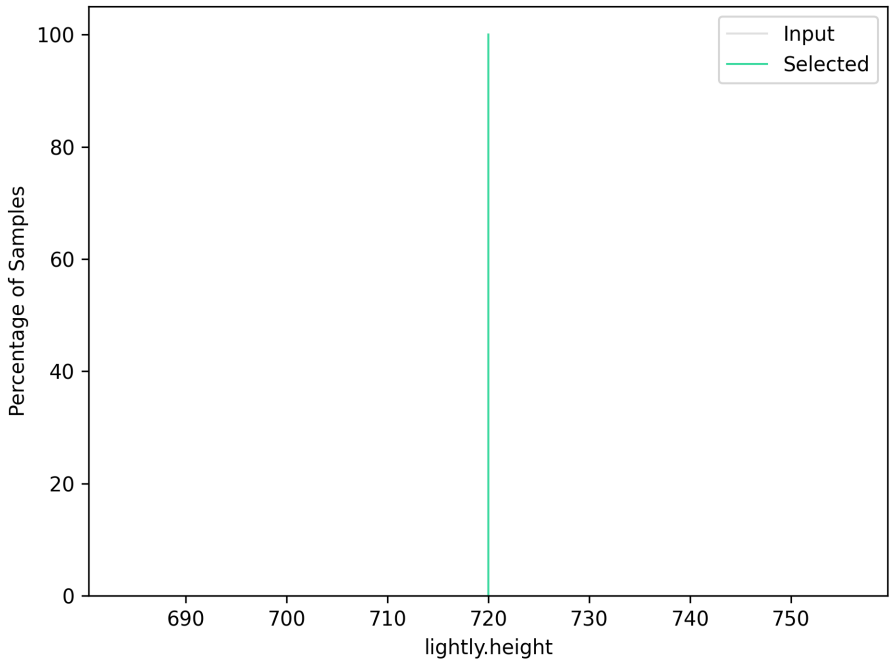
Retained (Left) and Removed (Right) Image with d = 0.00



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

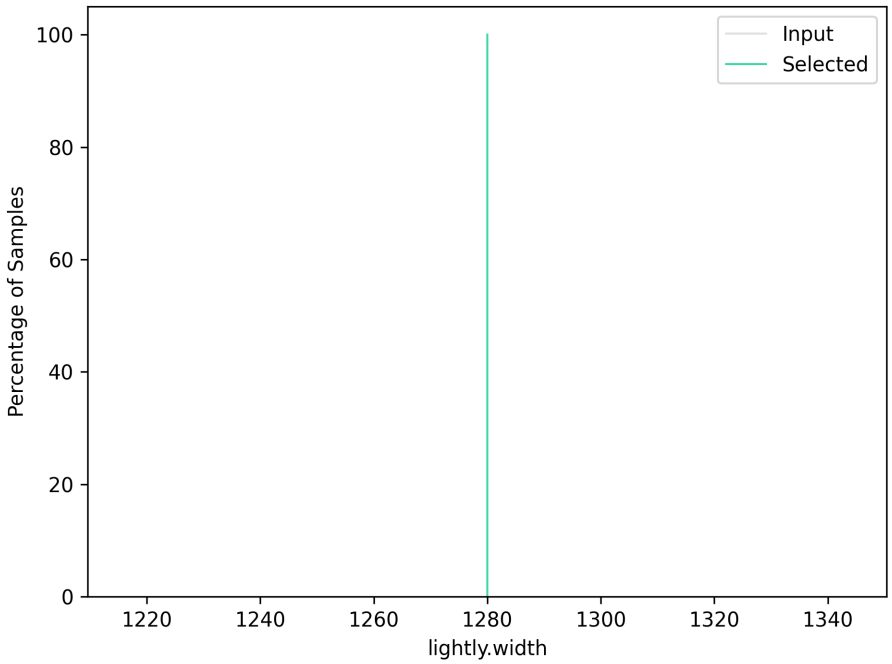
Input Height (lightly.height)



Set	Mean	Std	Min	Median	Max
Input	720	0	720	720	720
Selected	720	0	720	720	720

Lightly Metadata

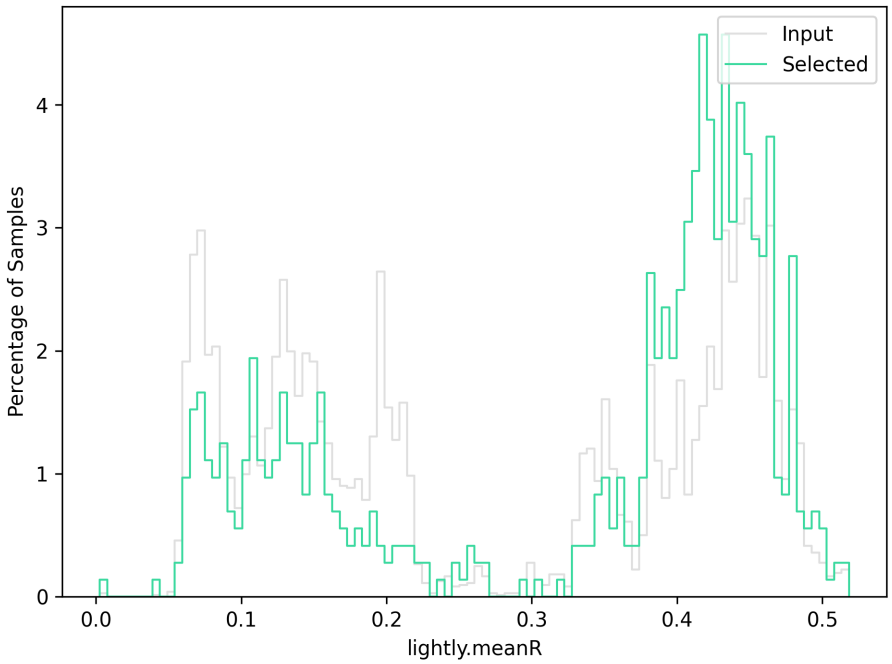
Input Width (lightly.width)



Set	Mean	Std	Min	Median	Max
Input	1280	0	1280	1280	1280
Selected	1280	0	1280	1280	1280

Lightly Metadata

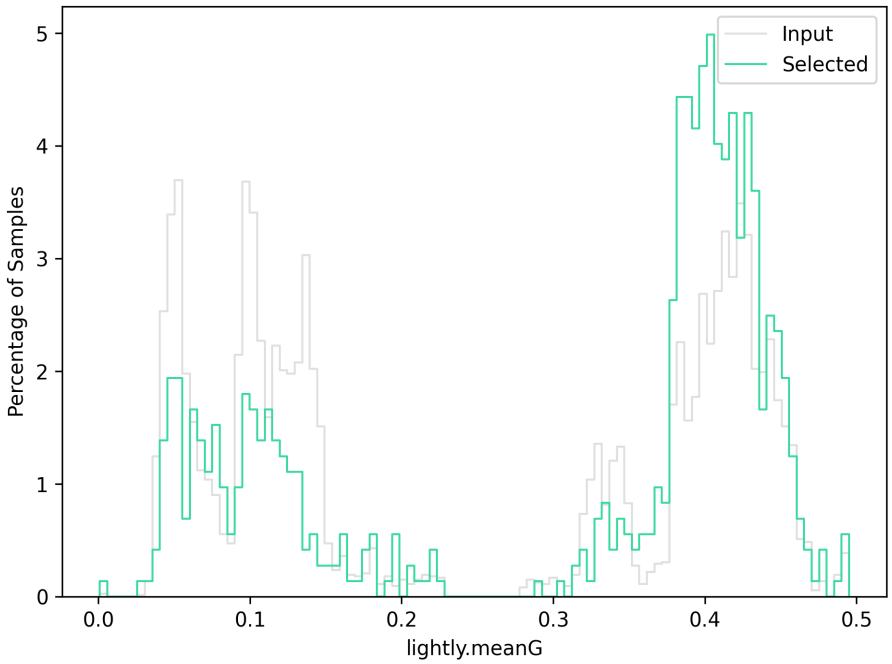
Red Channel Mean (lightly.meanR)



Set	Mean	Std	Min	Median	Max
Input	0.279	0.151	0.003	0.296	0.519
Selected	0.333	0.143	0.003	0.406	0.516

Lightly Metadata

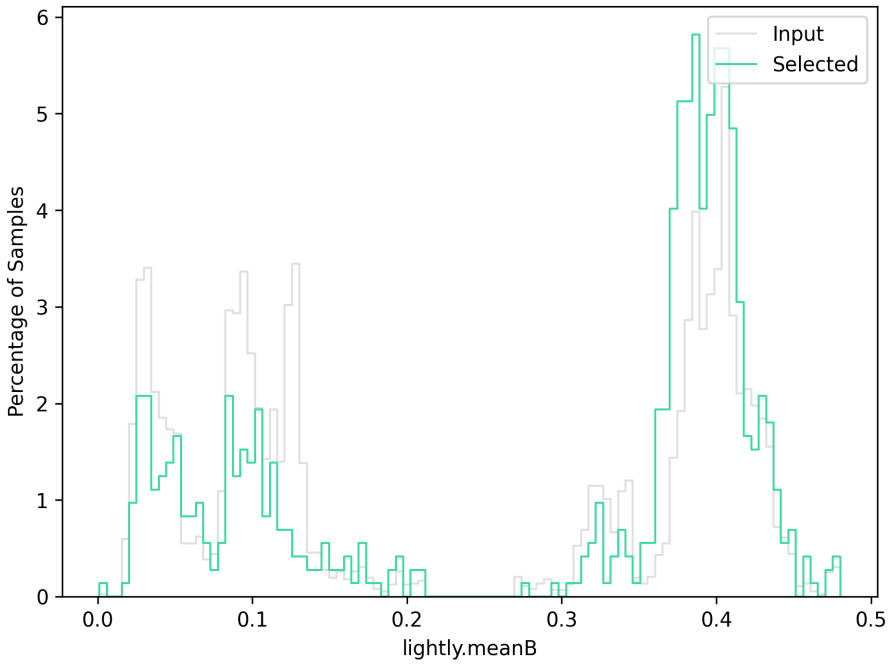
Green Channel Mean (lightly.meanG)



Set	Mean	Std	Min	Median	Max
Input	0.251	0.158	0.001	0.224	0.495
Selected	0.310	0.148	0.001	0.390	0.494

Lightly Metadata

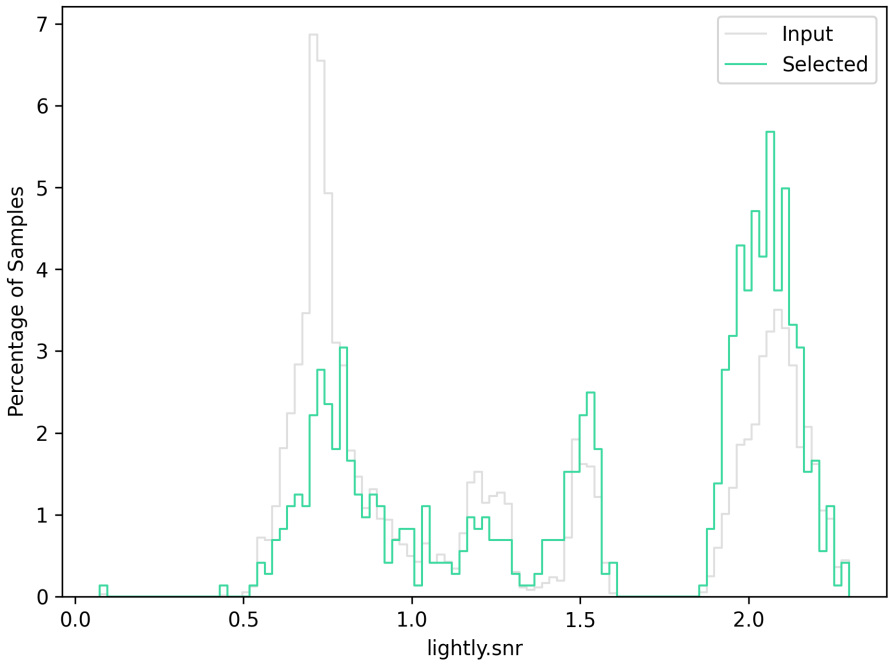
Blue Channel Mean (lightly.meanB)



Set	Mean	Std	Min	Median	Max
Input	0.237	0.158	0.001	0.210	0.480
Selected	0.296	0.148	0.001	0.379	0.480

Lightly Metadata

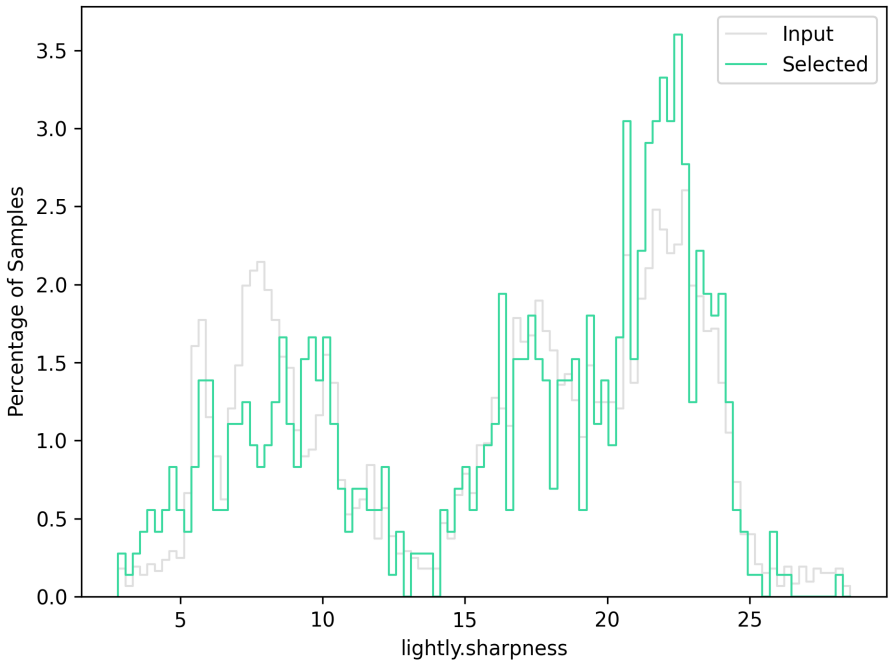
Signal to Noise Ratio (lightly.snr)



Set	Mean	Std	Min	Median	Max
Input	1.304	0.602	0.073	1.157	2.299
Selected	1.558	0.565	0.073	1.908	2.299

Lightly Metadata

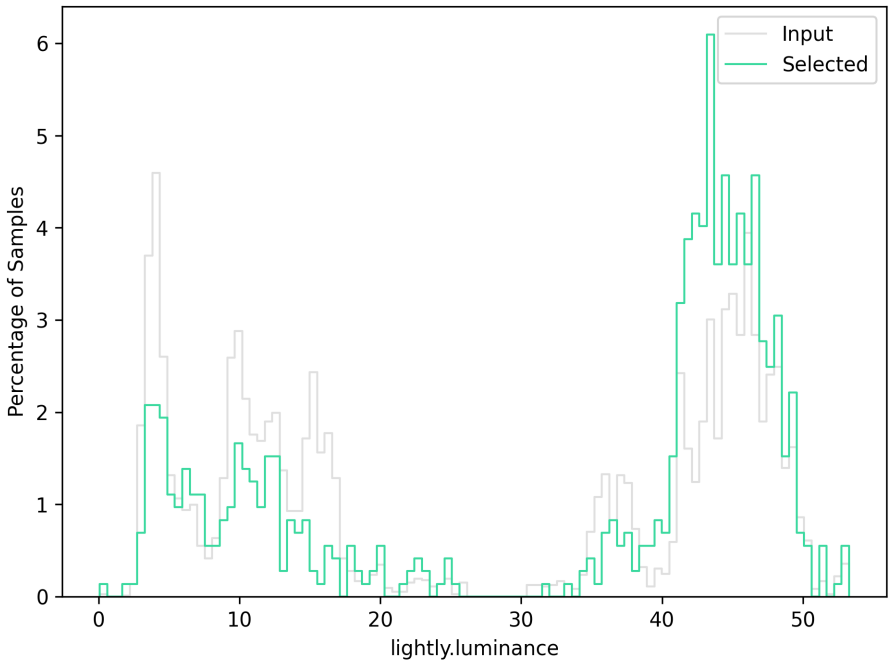
Sharpness (lightly.sharpness)



Set	Mean	Std	Min	Median	Max
Input	15.872	6.391	2.818	17.387	28.530
Selected	16.378	6.318	2.881	18.125	28.127

Lightly Metadata

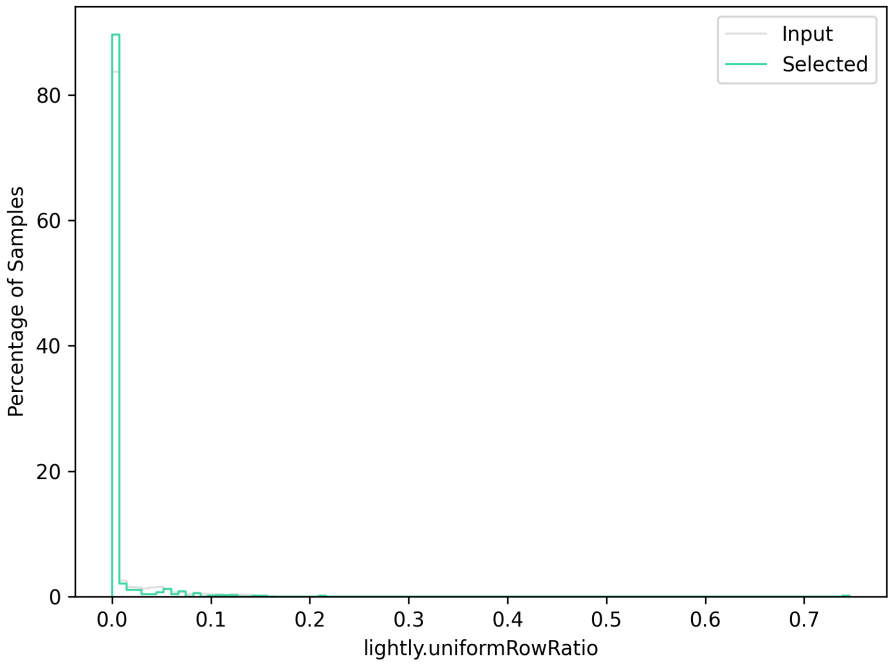
Luminance (lightly.luminance)



Set	Mean	Std	Min	Median	Max
Input	26.959	17.549	0.095	25.878	53.280
Selected	33.406	16.497	0.095	42.541	53.116

Lightly Metadata

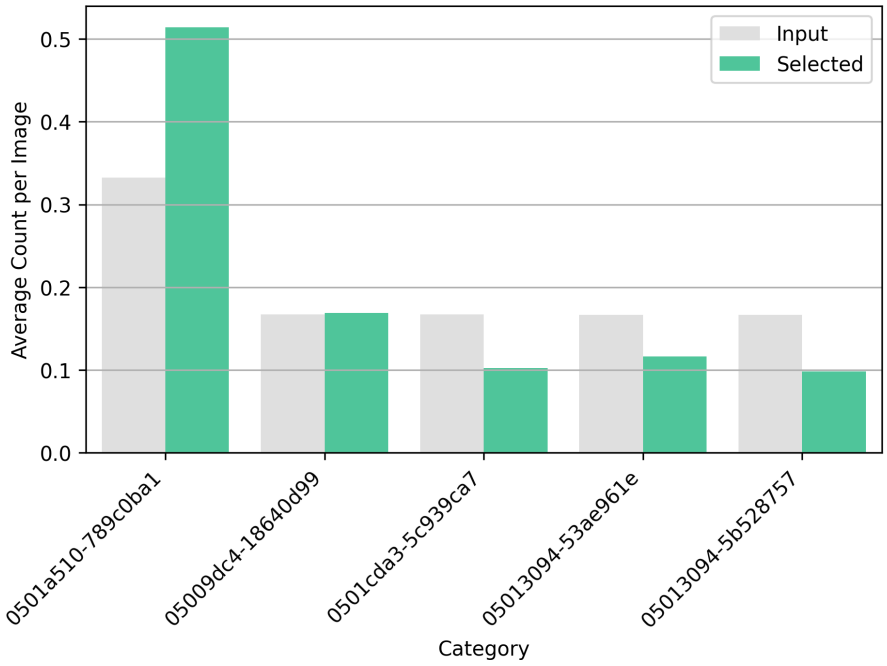
Uniform Row Ratio (lightly.uniformRowRatio)



Set	Mean	Std	Min	Median	Max
Input	0.009	0.027	0	0	0.746
Selected	0.006	0.034	0	0	0.746

Categorical Metadata: video_name

Video Name (video_name)



Average Category Counts per Image

Category	Input	Selected
05009dc4-18640d99	0.167	0.169
05013094-53ae961e	0.167	0.116
05013094-5b528757	0.167	0.098
0501a510-789c0ba1	0.332	0.514
0501cda3-5c939ca7	0.167	0.102
All Categories	1.000	1.000

Categorical Metadata: video_name

Category Distribution

Category	Input	Selected
05009dc4-18640d99	16.7%	16.9%
05013094-53ae961e	16.7%	11.6%
05013094-5b528757	16.7%	9.8%
0501a510-789c0ba1	33.2%	51.4%
0501cda3-5c939ca7	16.7%	10.2%
All Categories	100%	100%

Total Category Counts

Category	Input	Selected
05009dc4-18640d99	1208	122
05013094-53ae961e	1204	84
05013094-5b528757	1203	71
0501a510-789c0ba1	2401	371
0501cda3-5c939ca7	1207	74
All Categories	7223	722

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.