LIGHTLY

Dataset Filtering and Analytics Report

05/07/2023 13:15:19

LIGHTLY

General Information

Job Information

Metric	Value
Build Time	Wed Jul 5 12:55:11 UTC 2023
Build Version	2.7.dev
Job Submitted	05/07/2023 13:08:32
Job Finished	05/07/2023 13:15:19
Total Processing Time	06m 47s

Data Information

Metric	Images	Videos
Input	3608	2
Corrupt	0	0
Duplicates	0	N/A
Removed	3558	0
Output	50	2
Datapool Input	100	3
Datapool Output	150	5

Estimated Savings

Task	Annotation Savings*	CO2 Savings*
Image Classification	$ 1067.40	0.23 kg
Object Detection	$ 4269.60	0.85 kg
Semantic Segmentation	$ 21348.00	15.30 kg

*https://lightly.ai/report

LIGHTLY

Statistics

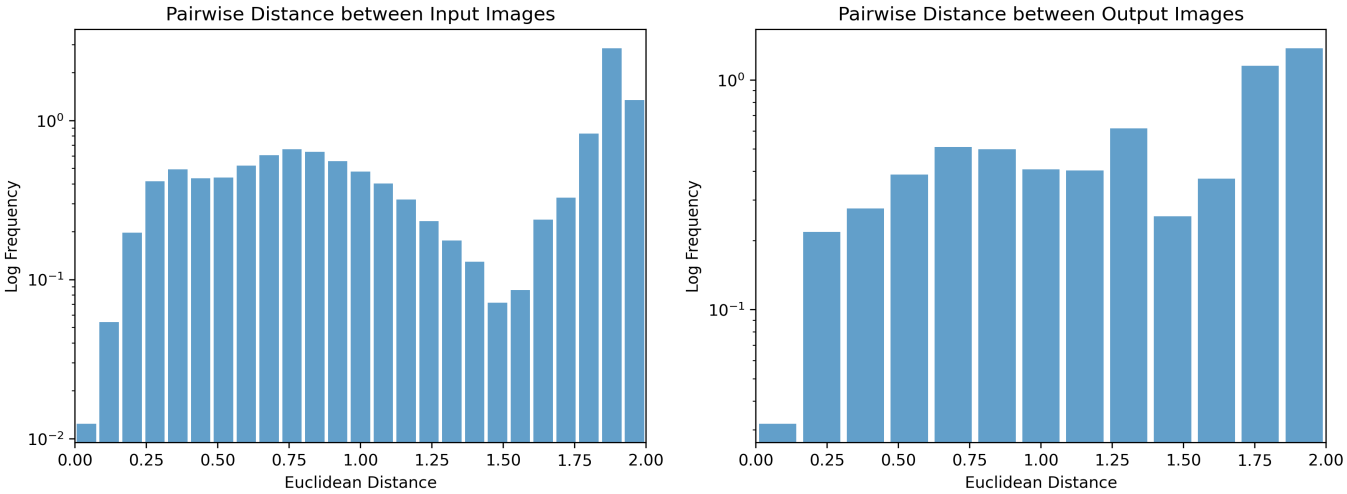
Distance

Metric	Before	After
Euclidean Distance (Mean)	1.2550	1.3015
Euclidean Distance (Min)	0.0026	0.0210
Euclidean Distance (Max)	1.9861	1.9728
Euclidean Distance (10th Percentile)	0.4133	0.5242
Euclidean Distance (90th Percentile)	1.9220	1.8963

Visualizations

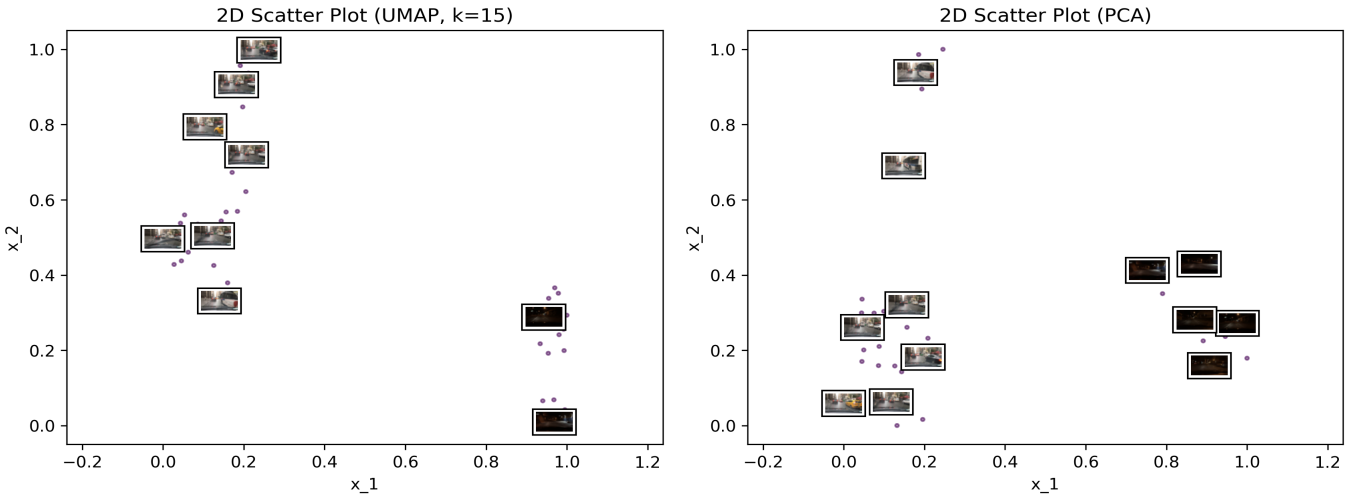
Image Similarity in Input and Output Data

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

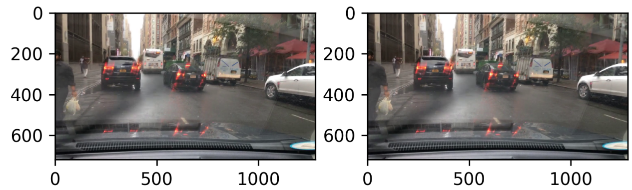
Two-dimensional scatter plots help to understand the distribution of the data and may enable quick insights about outlier cases, dataset bias, or class imbalances.



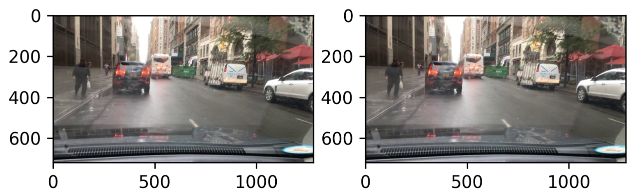
Sample of Retained Images and Similar Removed Images

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

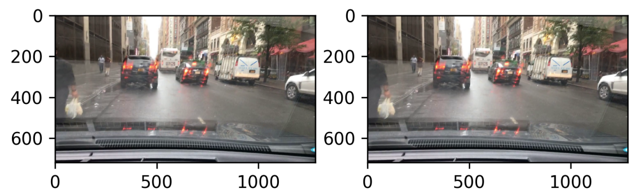
Retained (Left) and Removed (Right) Image with d = 0.00



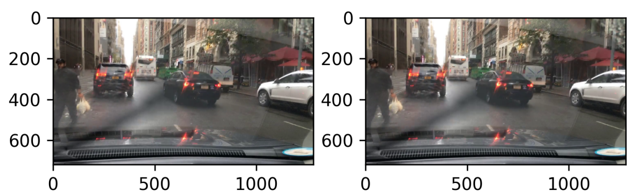
Retained (Left) and Removed (Right) Image with d = 0.00



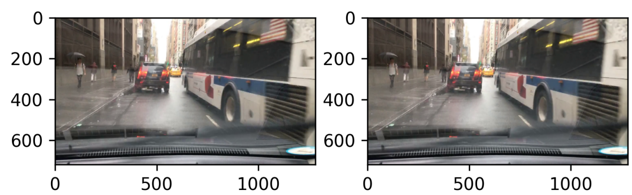
Retained (Left) and Removed (Right) Image with d = 0.00



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



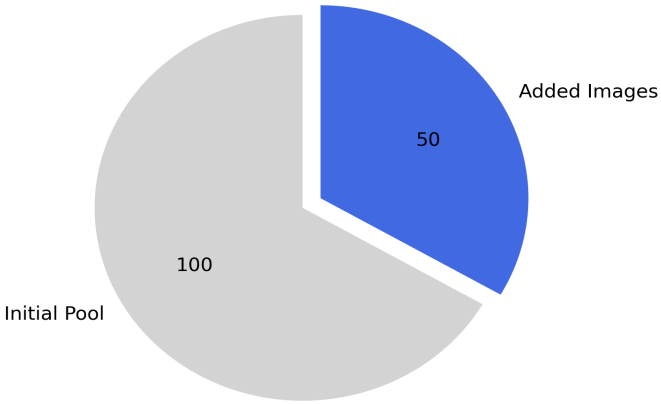
Retained (Left) and Removed (Right) Image with d = 0.00

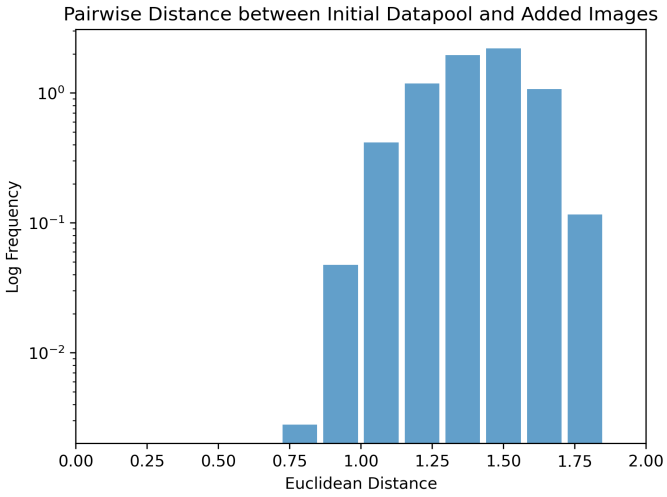


Datapool 1/2

Proportion of Added Images and Pairwise Distances

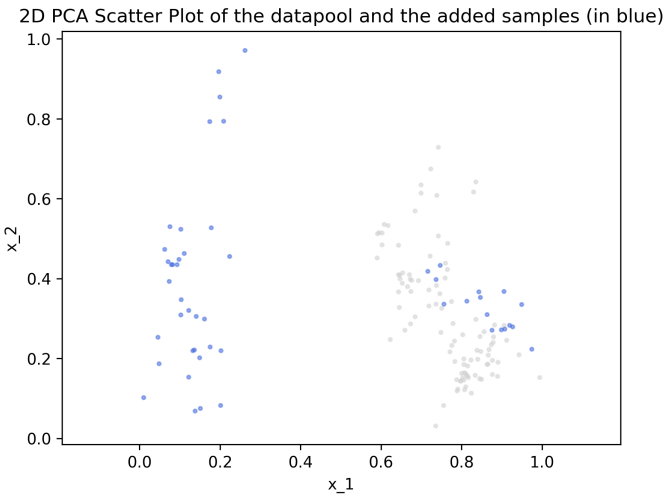
The figures below help to understand how many new images were added to the datapool and how similar the new images are to the ones selected in previous iterations.

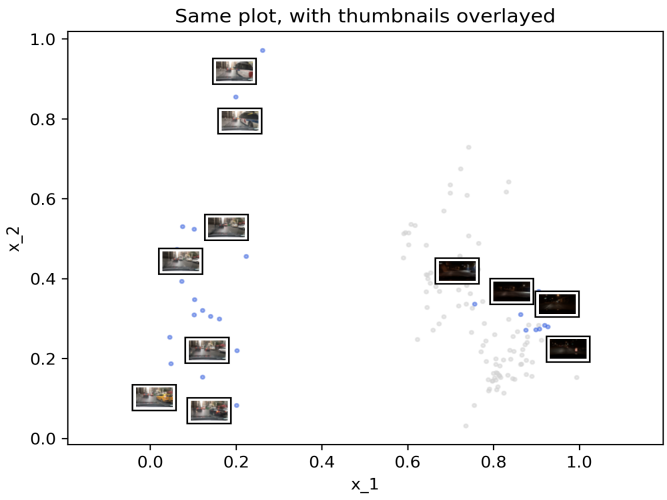




2D Scatter Plots of the Datapool (PCA)

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

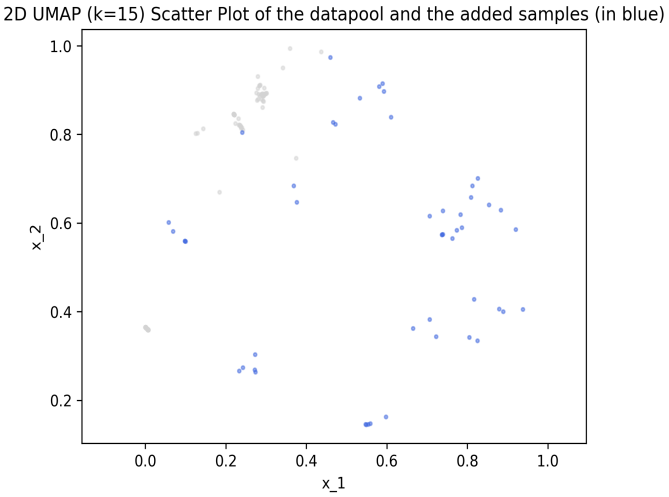


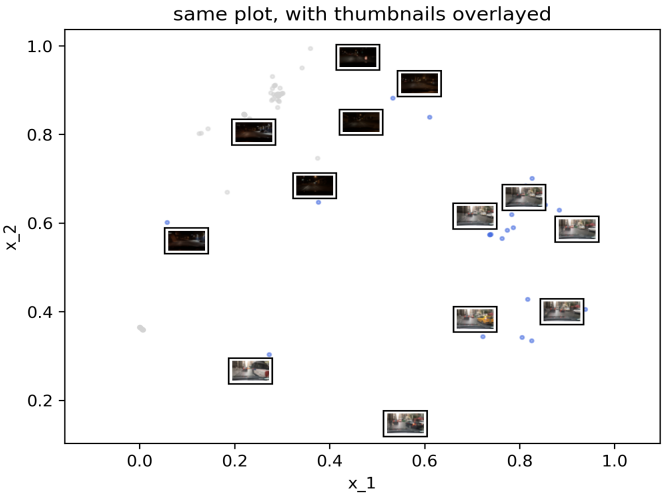


Datapool 2/2

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.

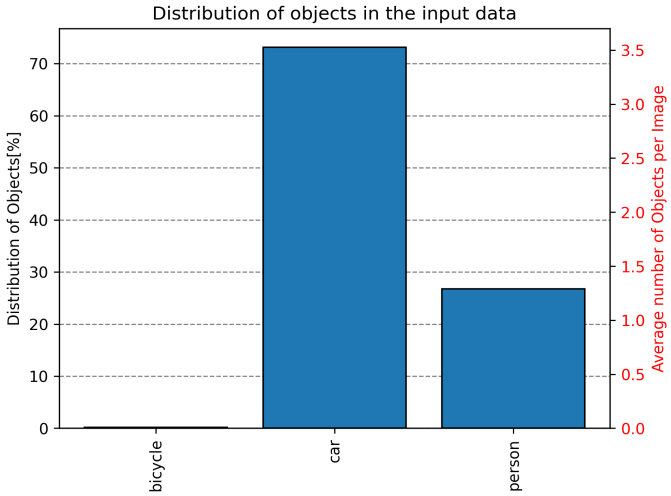


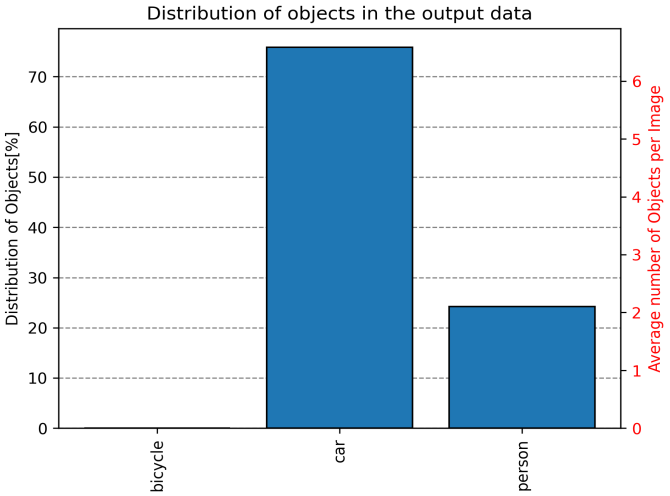


Prediction task: yolov8_detection

How did the distribution of predictions change?

Histogram plots of the number of objects found in the dataset.





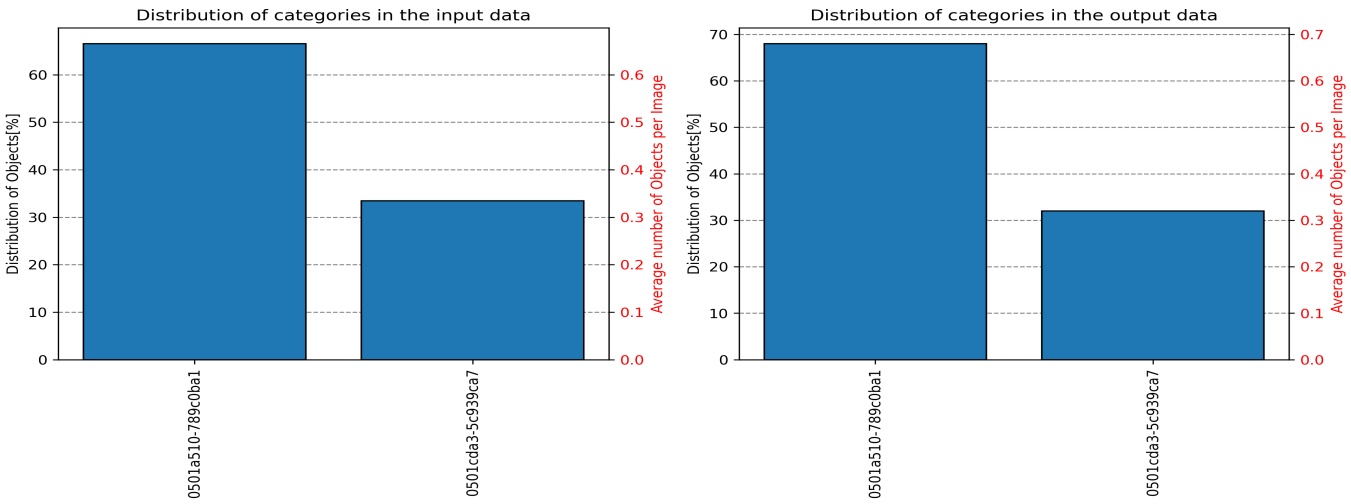
Total Objects

Object	Before Total	After Total	Before [%]	After [%]
bicycle	26.0	0.0	0.1	0.0
car	12716.0	329.0	73.1	75.8
person	4655.0	105.0	26.8	24.2

Metadata: video_name

How did the distribution of metadata categories change?

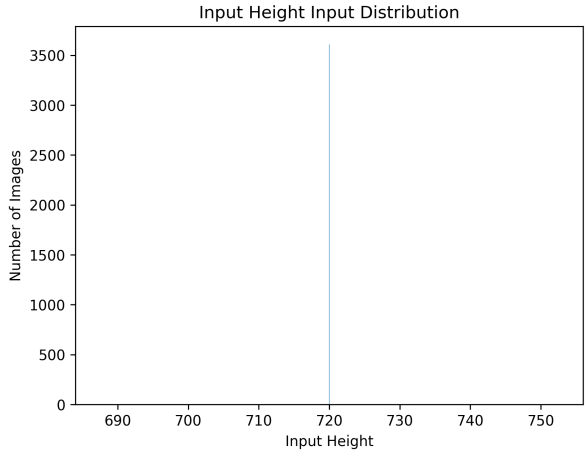
Histogram plots of the categories found in the dataset.

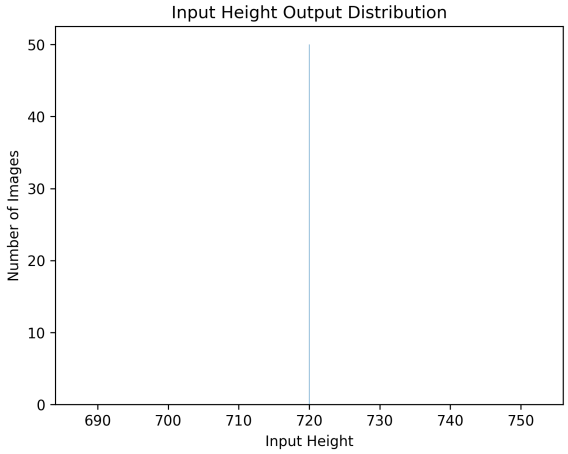


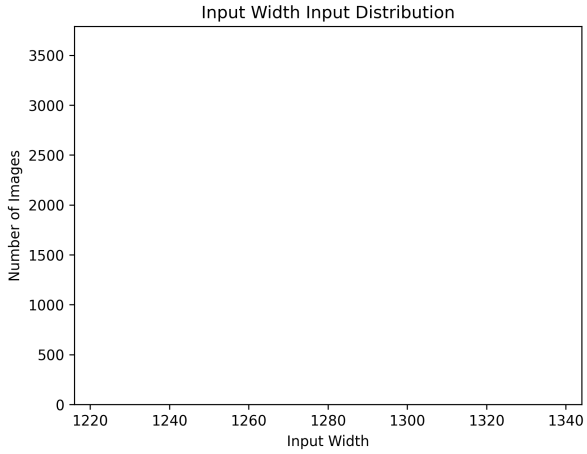
Total Categories

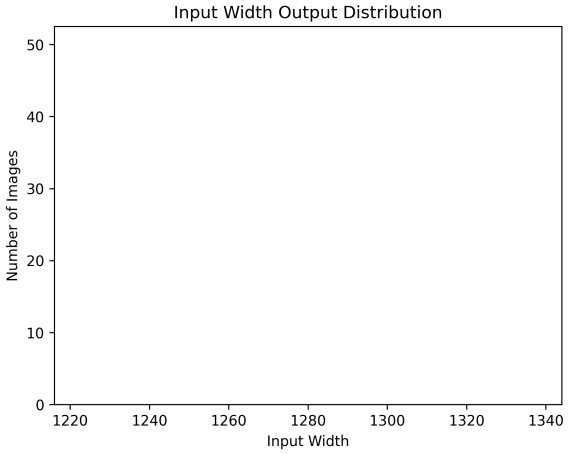
Category	Before Total	After Total	Before [%]	After [%]
0501a510-789c0ba1	2401	34	66.5	68.0
0501cda3-5c939ca7	1207	16	33.5	32.0

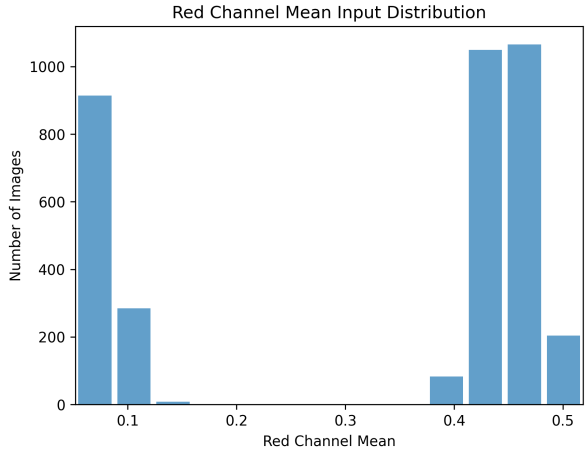
Metadata

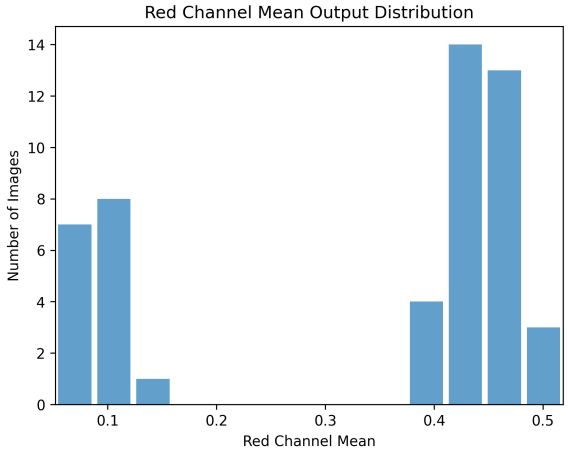

Input Height Input Distribution
Number of Images
Input Height
0
500
1000
1500
2000
2500
3000
3500
690
700
710
720
730
740
750


Input Height Output Distribution
Number of Images
Input Height
0
10
20
30
40
50
690
700
710
720
730
740
750

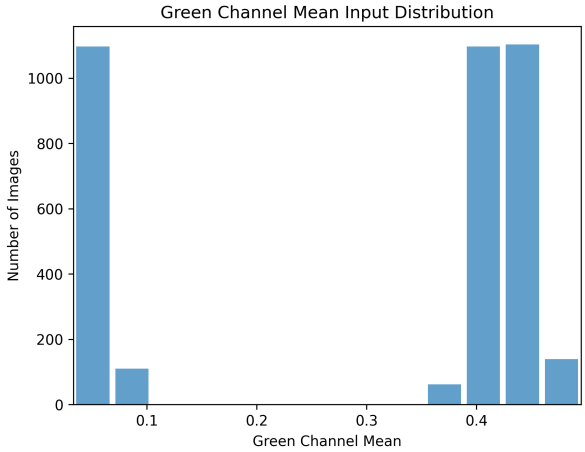

Input Width Input Distribution
Number of Images
Input Width
0
500
1000
1500
2000
2500
3000
3500
1220
1240
1260
1280
1300
1320
1340

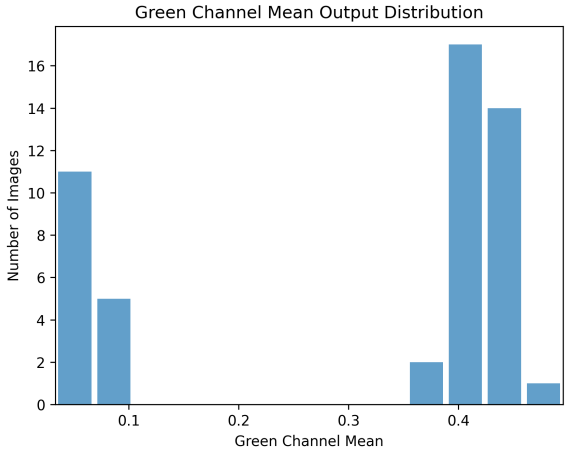

Input Width Output Distribution
Number of Images
Input Width
0
10
20
30
40
50
1220
1240
1260
1280
1300
1320
1340

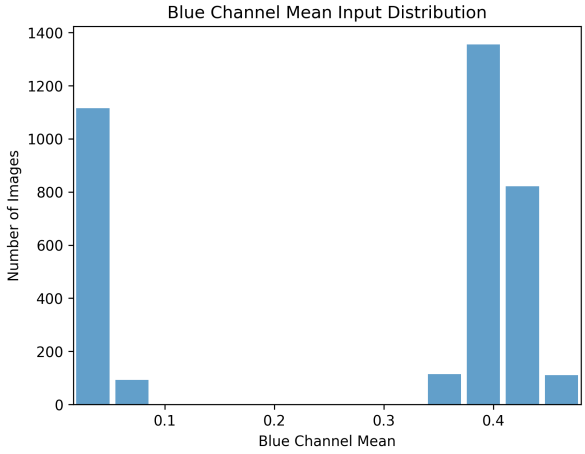

Red Channel Mean Input Distribution
Number of Images
Red Channel Mean
0
200
400
600
800
1000
0.1
0.2
0.3
0.4
0.5

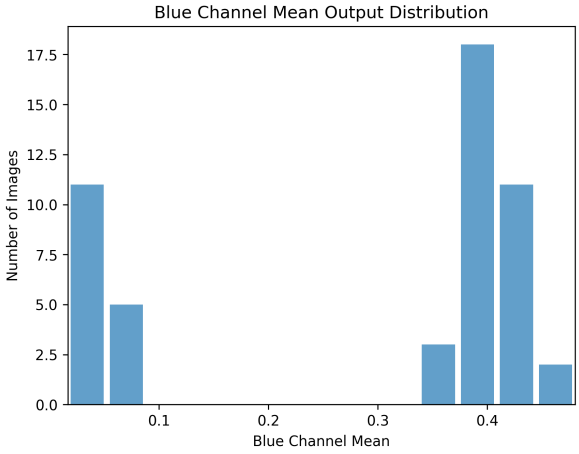

Red Channel Mean Output Distribution
Number of Images
Red Channel Mean
0
2
4
6
8
10
12
14
0.1
0.2
0.3
0.4
0.5

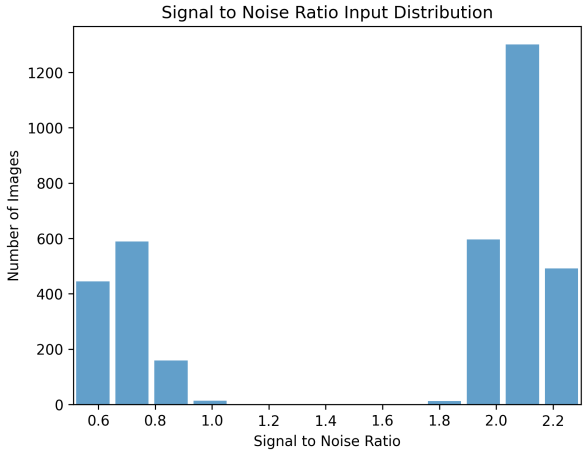
Metadata

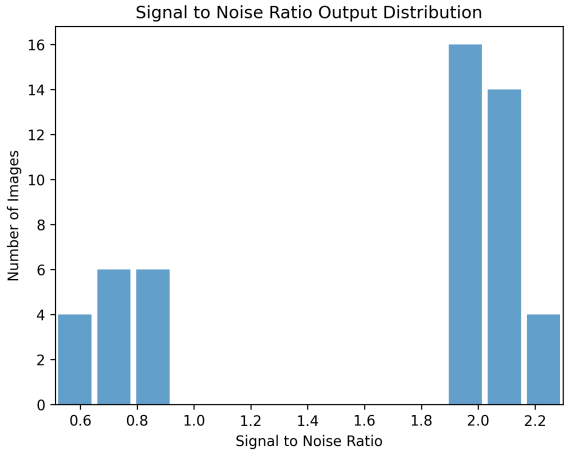

Green Channel Mean Input Distribution
Number of Images
Green Channel Mean


Green Channel Mean Output Distribution
Number of Images
Green Channel Mean

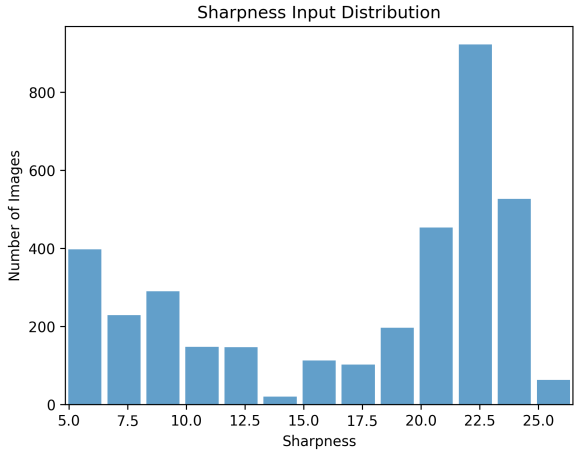

Blue Channel Mean Input Distribution
Number of Images
Blue Channel Mean

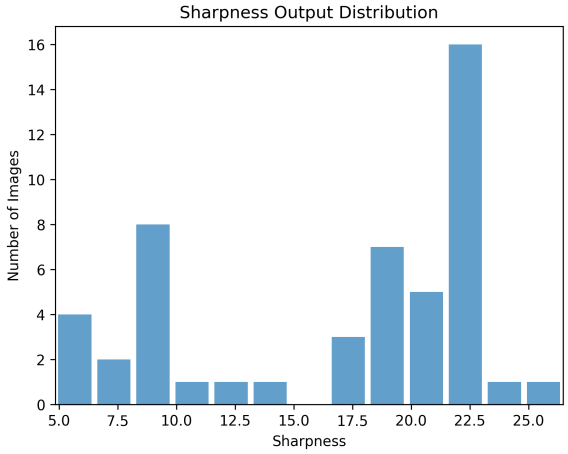

Blue Channel Mean Output Distribution
Number of Images
Blue Channel Mean

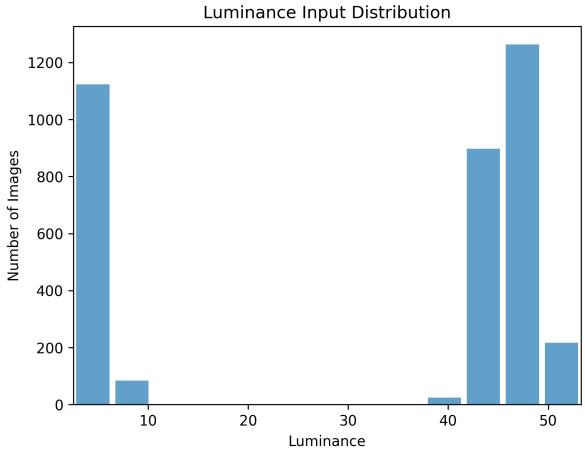

Signal to Noise Ratio Input Distribution
Number of Images
Signal to Noise Ratio

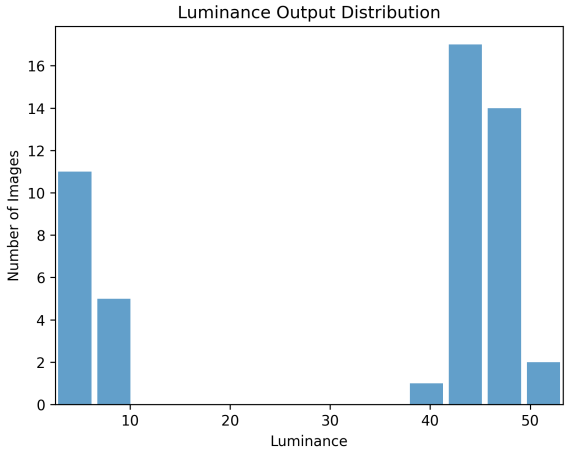

Signal to Noise Ratio Output Distribution
Number of Images
Signal to Noise Ratio

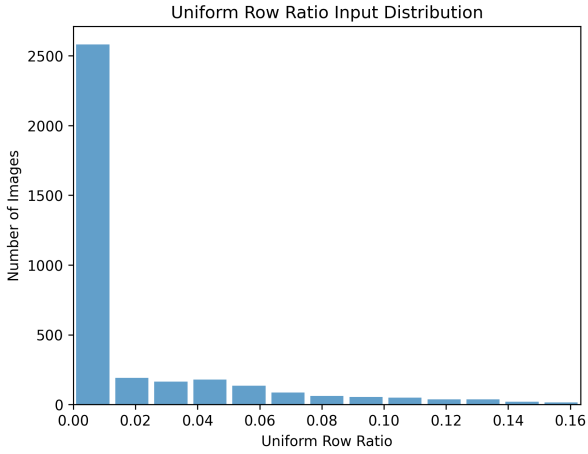
Metadata

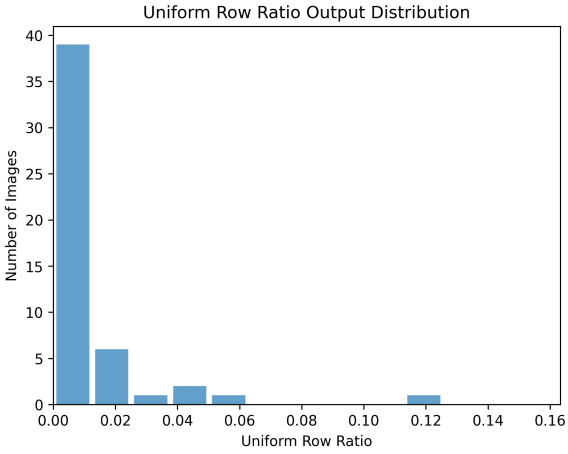

Sharpness Input Distribution
Number of Images
Sharpness


Sharpness Output Distribution
Number of Images
Sharpness


Luminance Input Distribution
Number of Images
Luminance


Luminance Output Distribution
Number of Images
Luminance


Uniform Row Ratio Input Distribution
Number of Images
Uniform Row Ratio


Uniform Row Ratio Output Distribution
Number of Images
Uniform Row Ratio

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.