



# LIGHTLY

## Dataset Filtering and Analytics Report

05/07/2023 13:07:52

# 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 12:58:32
Job Finished	05/07/2023 13:07:52
Total Processing Time	09m 20s

## Data Information

Metric	Images	Videos
Input	3615	3
Corrupt	0	0
Duplicates	0	N/A
Removed	3515	0
Output	100	3
Datapool Input	0	0
Datapool Output	100	3

## Estimated Savings

Task	Annotation Savings*	CO2 Savings* 
Image Classification	\$ 1054.50	0.23 kg
Object Detection	\$ 4218.00	0.84 kg
Semantic Segmentation	\$ 21090.00	15.11 kg

\*<https://lightly.ai/report>

# Statistics

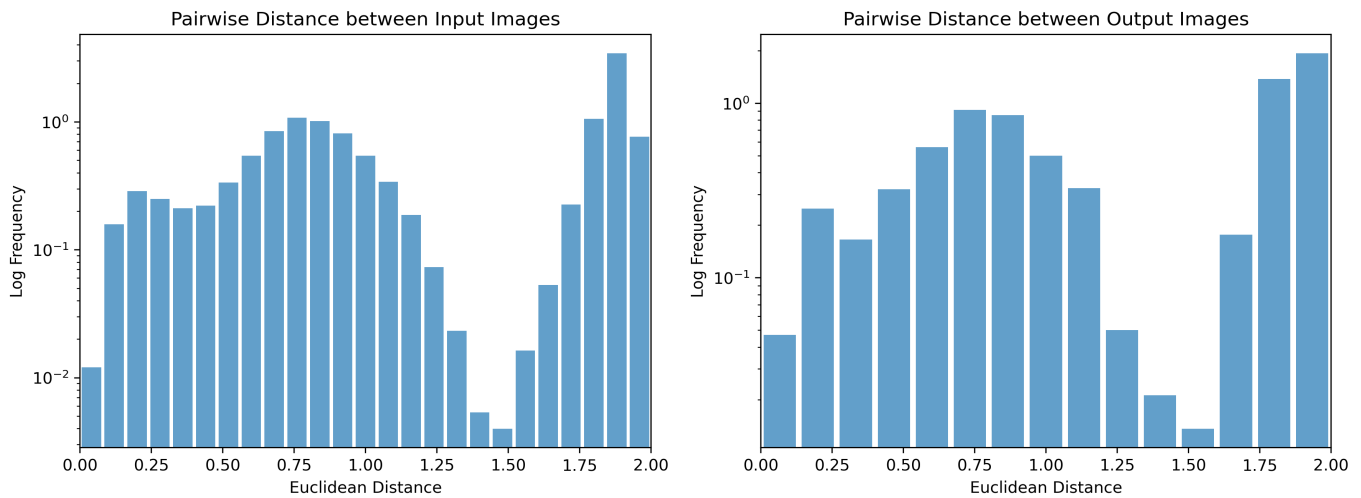
## Distance

Metric	Before	After
Euclidean Distance (Mean)	1.2361	1.2626
Euclidean Distance (Min)	0.0020	0.0587
Euclidean Distance (Max)	1.9916	1.9797
Euclidean Distance (10th Percentile)	0.5117	0.5240
Euclidean Distance (90th Percentile)	1.9084	1.9167

# Visualizations

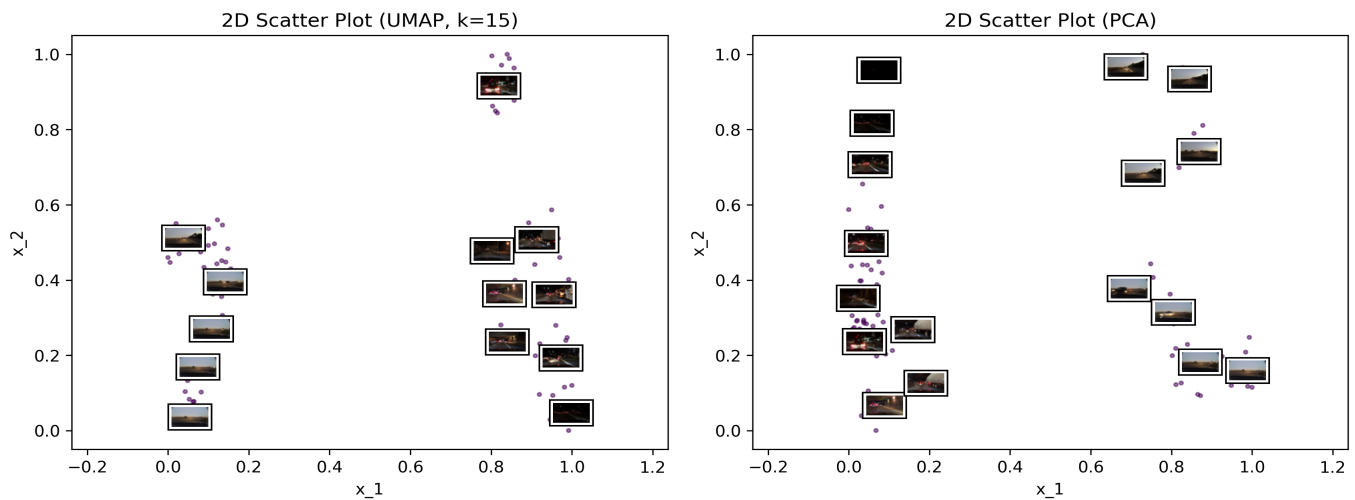
## Image Similarity in Input and Output Data

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.



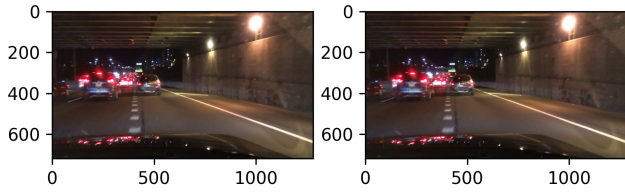
## 2D Scatter Plots of Output Data

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.

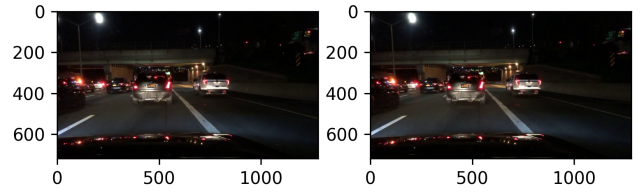


# Sample of Retained Images and Similar Removed Images

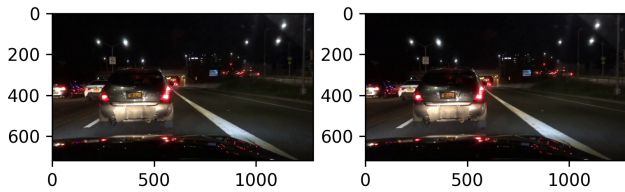
Retained (Left) and Removed (Right) Image with  $d = 0.00$



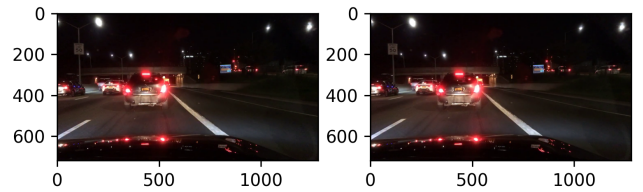
Retained (Left) and Removed (Right) Image with  $d = 0.01$



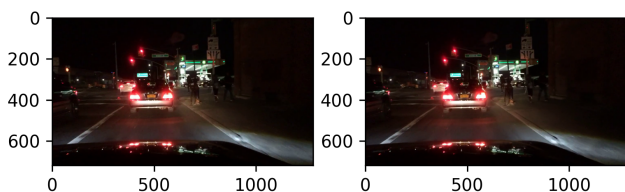
Retained (Left) and Removed (Right) Image with  $d = 0.01$



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



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



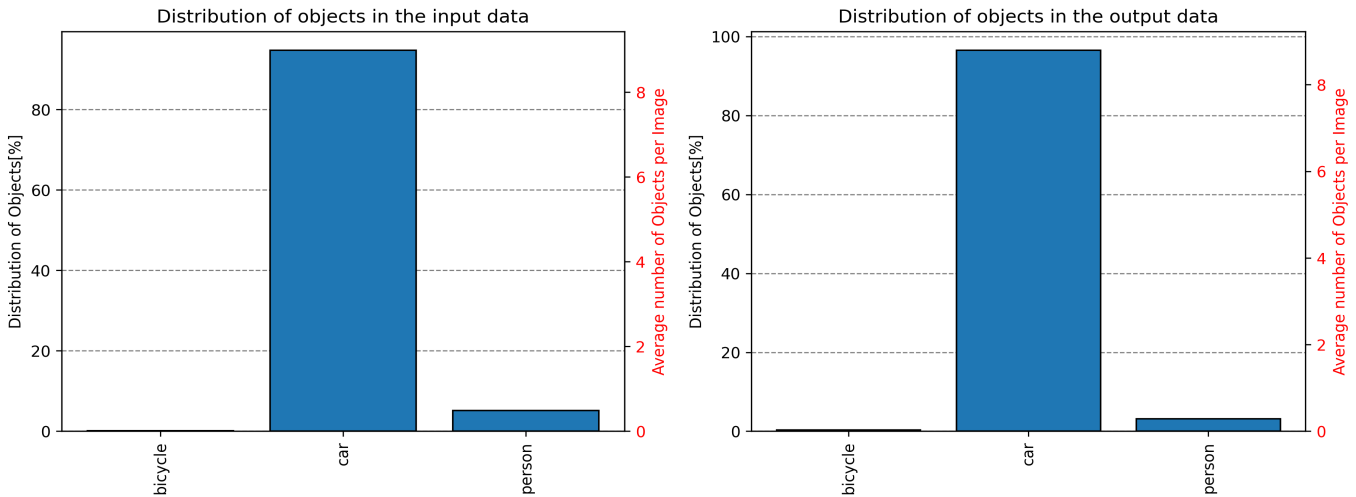
Retained (Left) and Removed (Right) Image with  $d = 0.01$



# Prediction task: yolov8\_detection

## How did the distribution of predictions change?

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



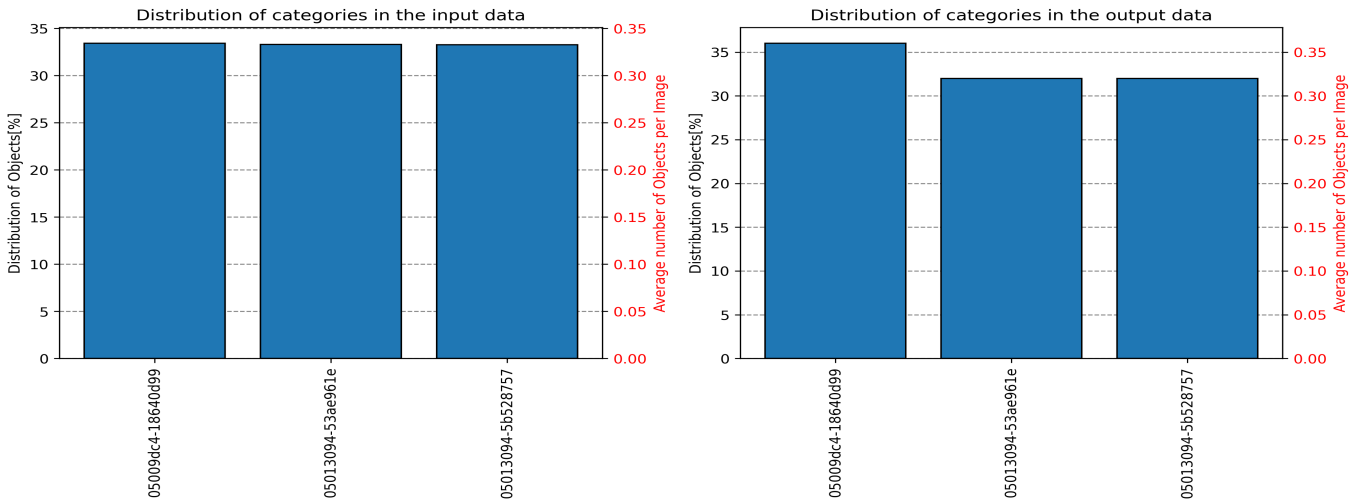
## Total Objects

Object	Before Total	After Total	Before [%]	After [%]
bicycle	45.0	3.0	0.1	0.3
car	32484.0	879.0	94.7	96.5
person	1769.0	29.0	5.2	3.2

# Metadata: video\_name

## How did the distribution of metadata categories change?

Histogram plots of the categories found in the dataset.

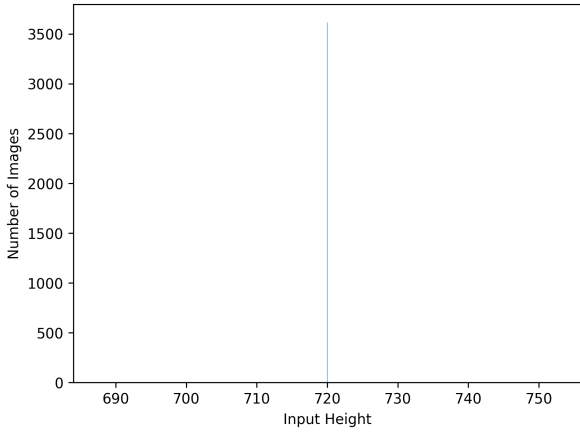


## Total Categories

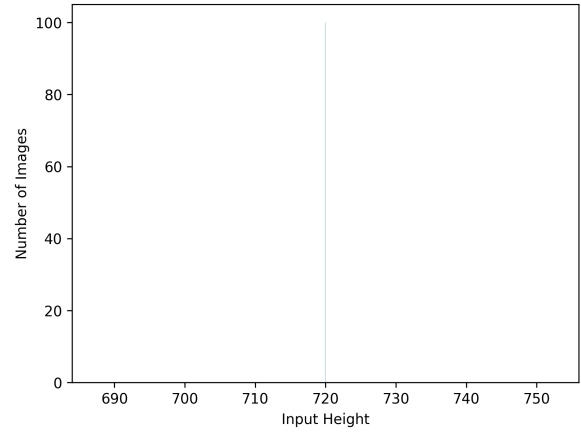
Category	Before Total	After Total	Before [%]	After [%]
05009dc4-18640d99	1208	36	33.4	36.0
05013094-53ae961e	1204	32	33.3	32.0
05013094-5b528757	1203	32	33.3	32.0

# Metadata

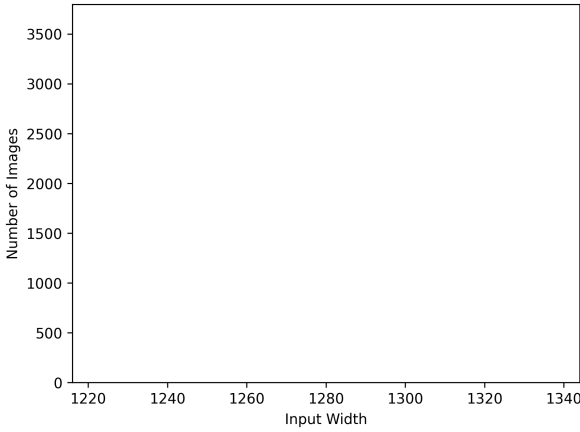
Input Height Input Distribution



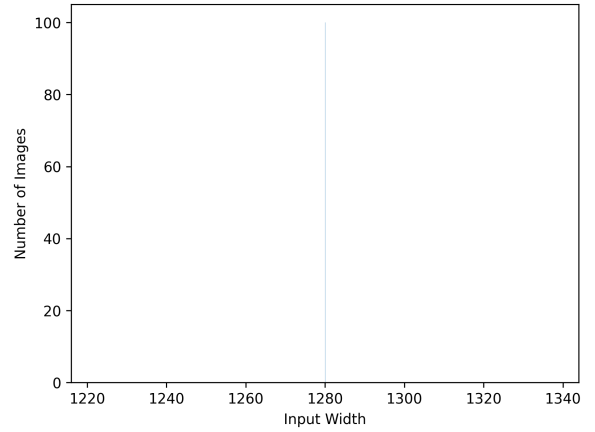
Input Height Output Distribution



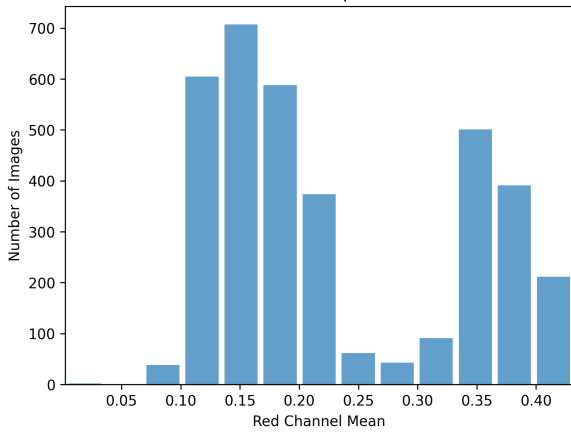
Input Width Input Distribution



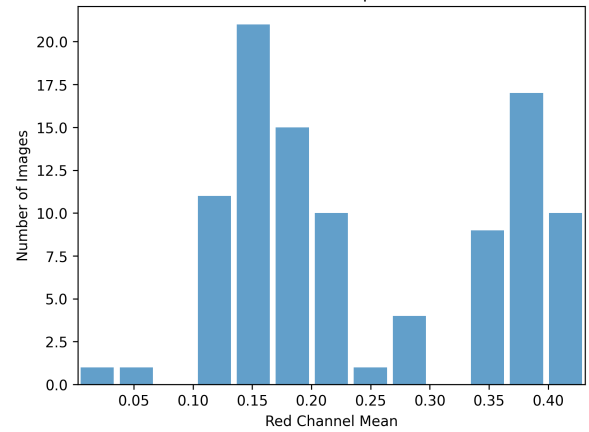
Input Width Output Distribution



Red Channel Mean Input Distribution

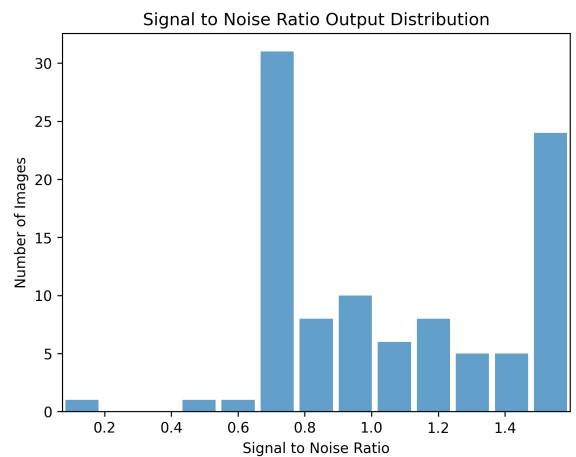
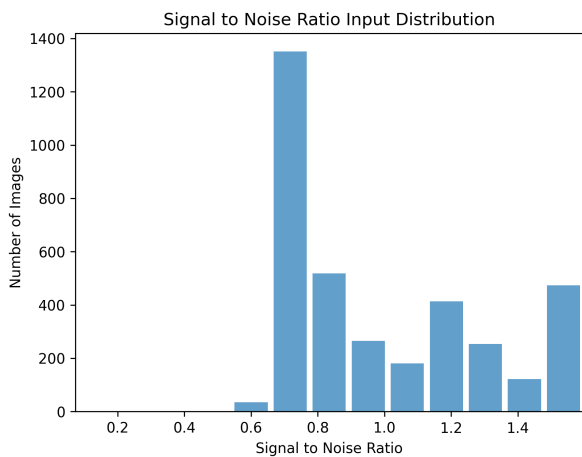
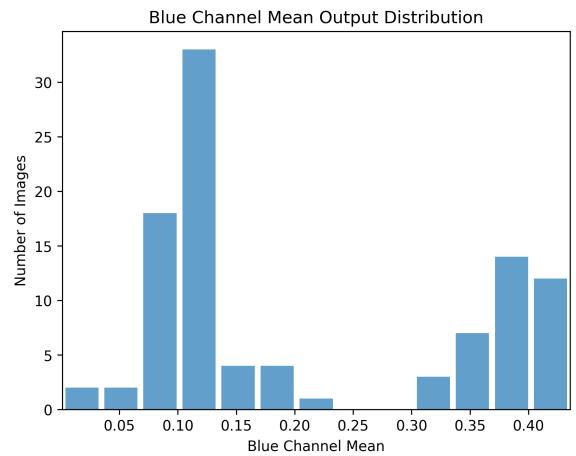
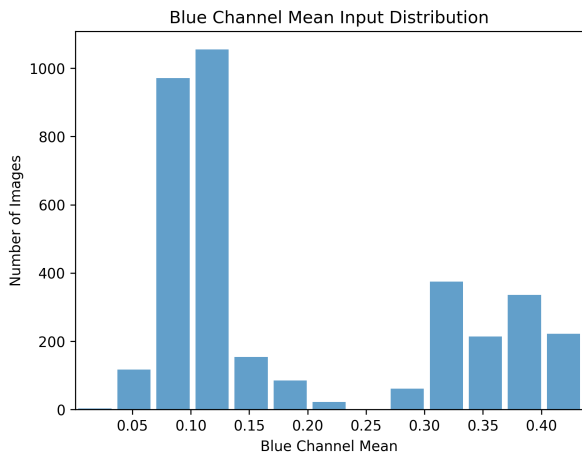
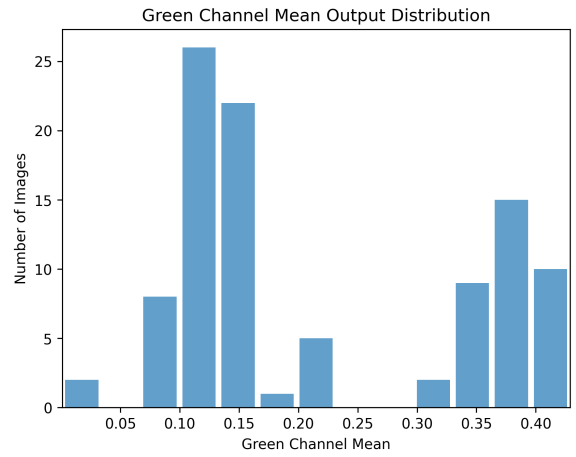
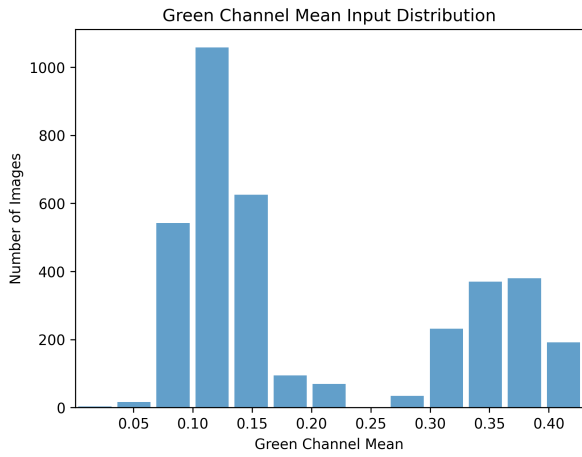


Red Channel Mean Output Distribution

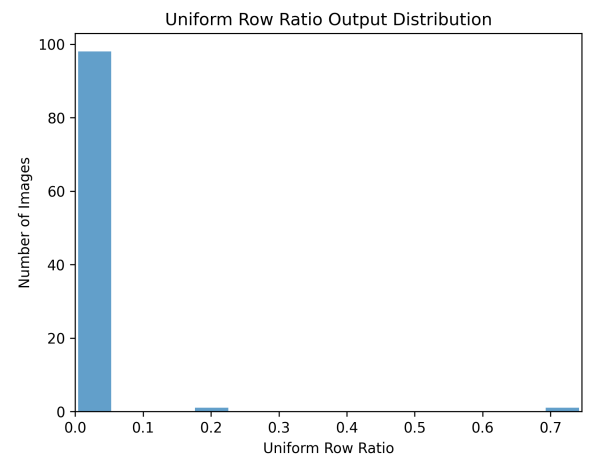
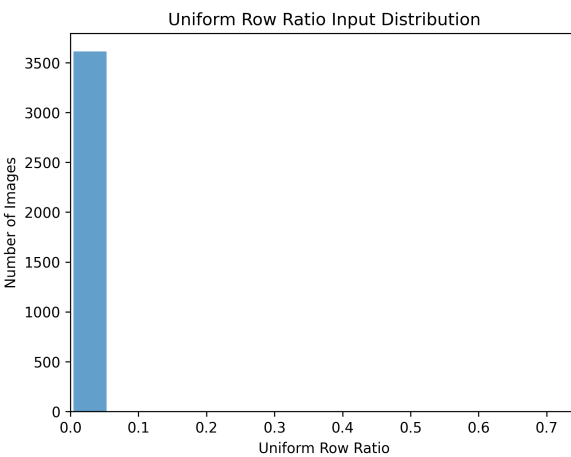
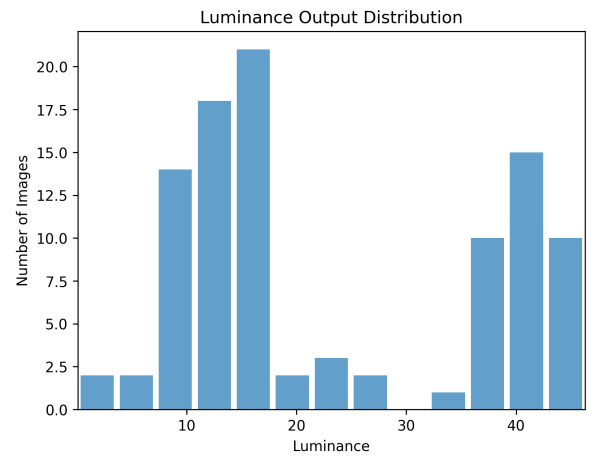
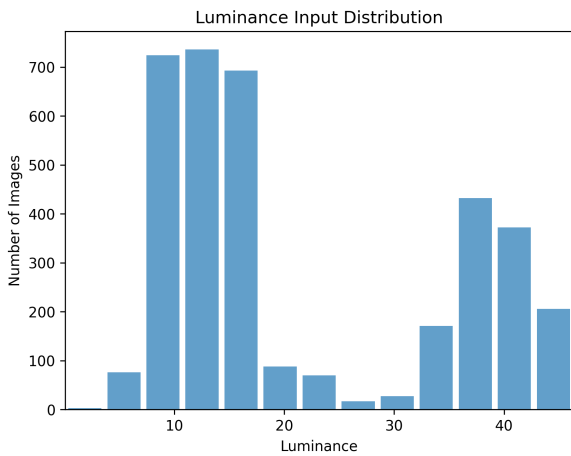
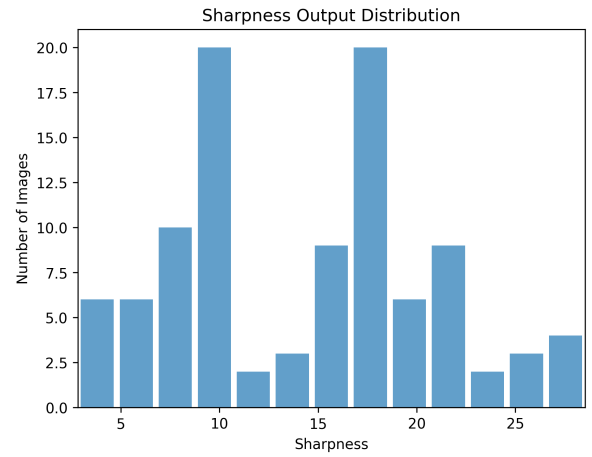
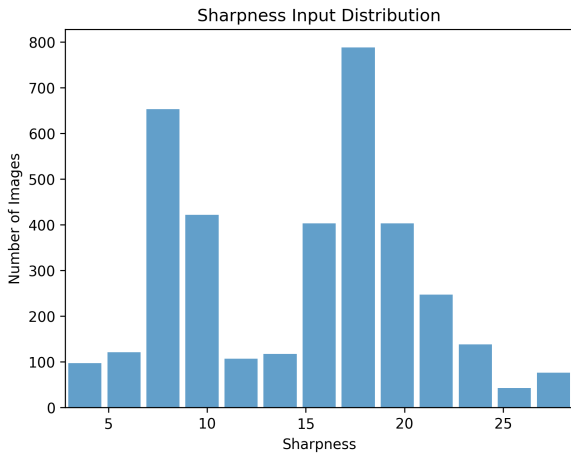




# Metadata



# Metadata



# Video Sampling Densities 1/1

We show selected frames for each video. Each selected frame is indicated by a vertical line.

When using coresets, clusters of selected frames show sequences where the frames differ a lot visually. Additionally, high density regions will appear darker in the plots.

