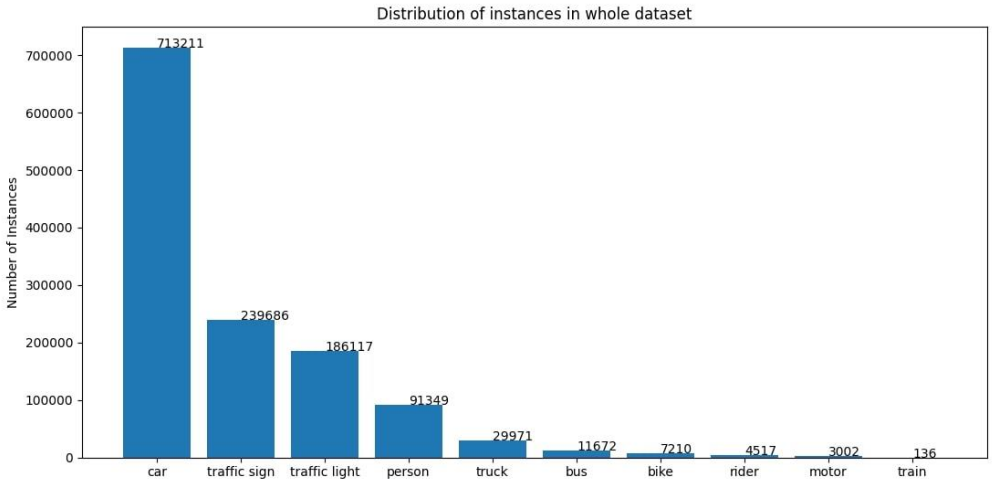
Exploratory Data Analysis on BDD100K Dataset

A comprehensive data analysis on the annotated BDD100K dataset to evaluate its suitability for object detection. This analysis reveals significant class imbalance, not only in the number of instances but also in other categorical features. To visualize these findings and deeper insights. Below are the plots:

Analysis on Training Set

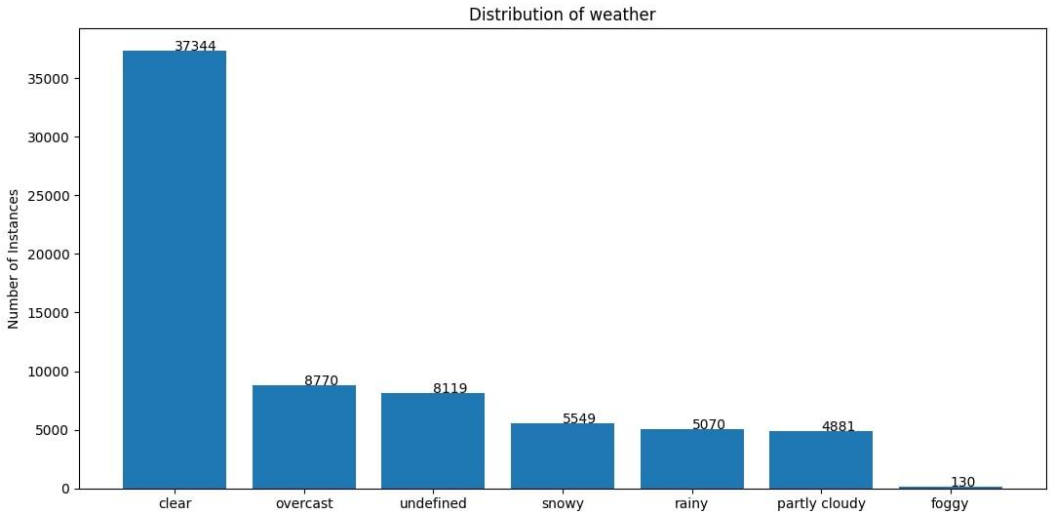
1. **Distribution of instances**



This bar chart depicts the distribution of instances in a dataset. Here's an evaluation of how this distribution impacts model training:

- "Car" dominates the dataset with over 700,000 instances.
- Rare classes like "train" (136) and "motor" (3,002) are severely underrepresented.
- The model might focus too much on cars and ignore smaller classes like "train" or "rider."
- Classes with fewer samples (e.g., "motor," "bike," "train") will likely have poor detection accuracy.

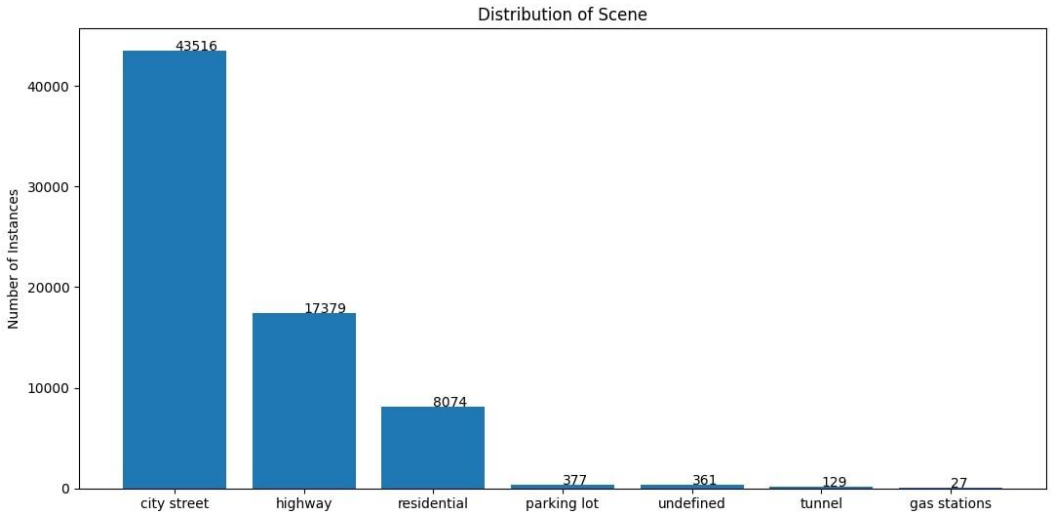
2. Distribution of weather images



This bar chart depicts the distribution of instances in a dataset. Here's an evaluation of how this distribution impacts model training:

- **Clear weather dominates** with most data, so the model might focus more on clear weather and ignore others.
- **Foggy weather is rare** (only 130 images), making it hard for the model to learn in such conditions.
- Other weather types like **rainy and snowy** have less examples, so the model may underperform in those scenarios.

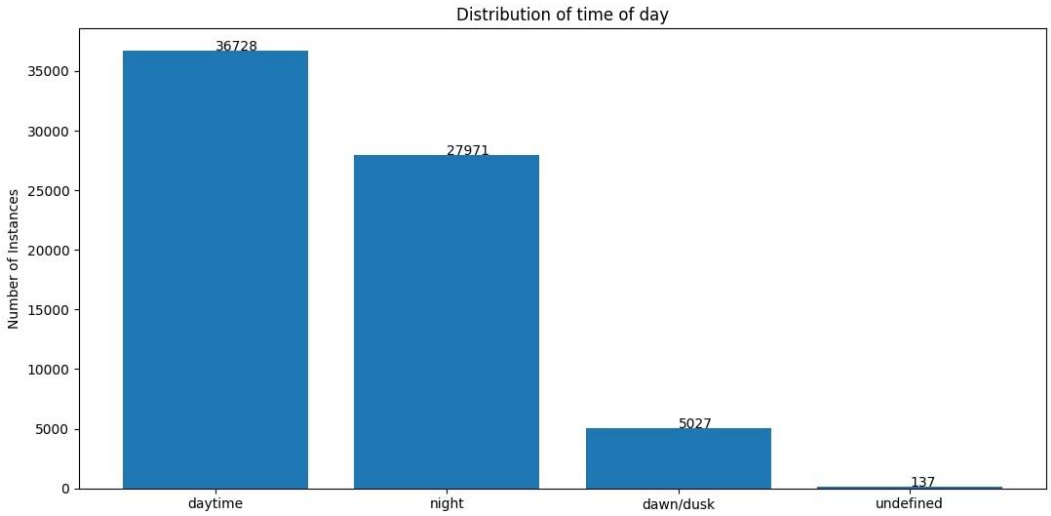
3. Distribution of various scene images



This bar chart depicts the distribution of instances in a dataset. Here's an evaluation of how this distribution impacts model training:

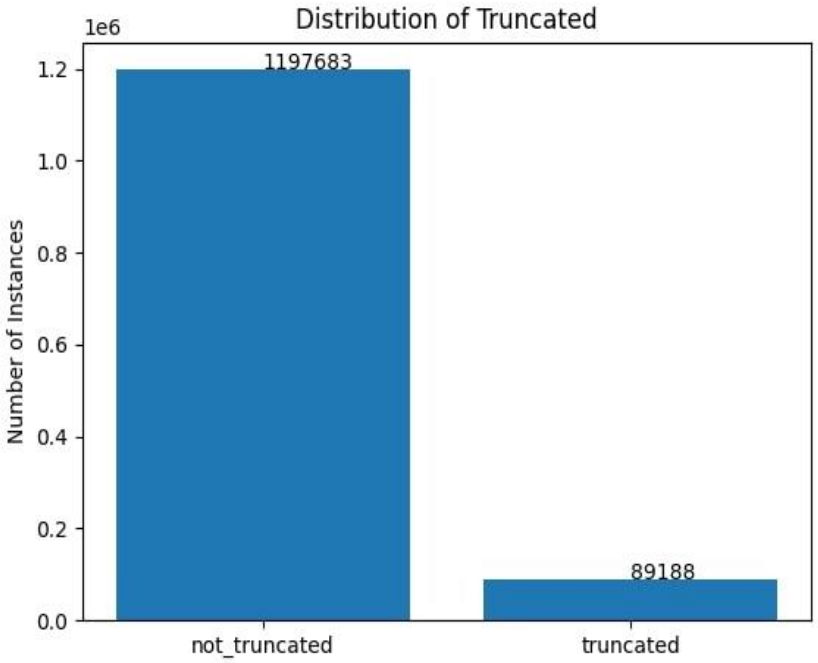
- The dataset is heavily skewed towards the "city street" class, which could result into biased model.
- The presence of an "undefined" class indicates potential data ambiguity or mislabeling.
- While the overall dataset size seems reasonable, some classes have limited instances, which might affect model performance.

4. Distribution of day/night images



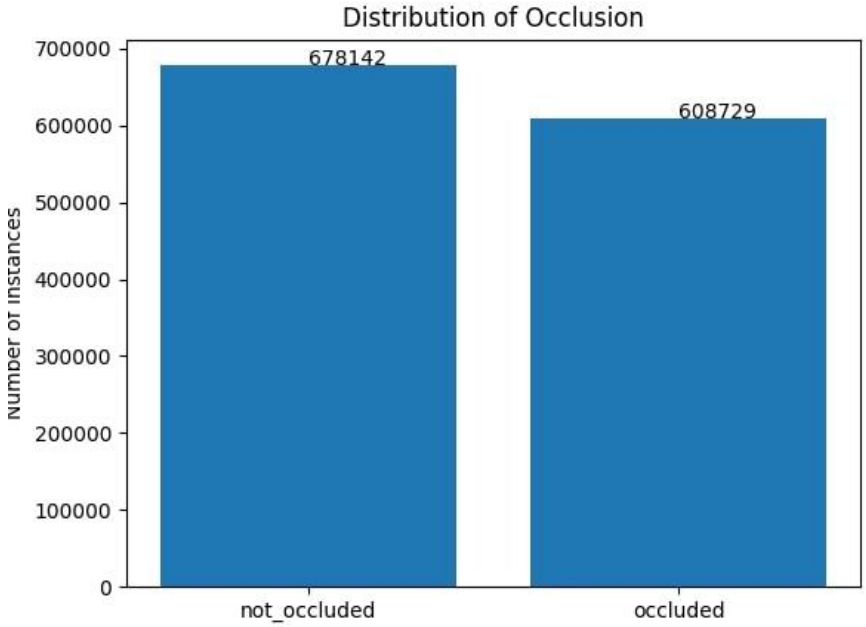
- The dataset is heavily skewed towards the "daytime" class, which might bias the model towards daytime scenarios and hinder its performance in low-light conditions.
- The presence of an "undefined" class indicates potential data ambiguity or mislabeling, which can introduce noise and negatively impact the model's training.
- While the overall dataset size seems reasonable, the "dawn/dusk" and "undefined" classes have significantly fewer instances, potentially limiting the model's ability to learn and generalize in those specific conditions.

5. **Instances of Truncated annotations**



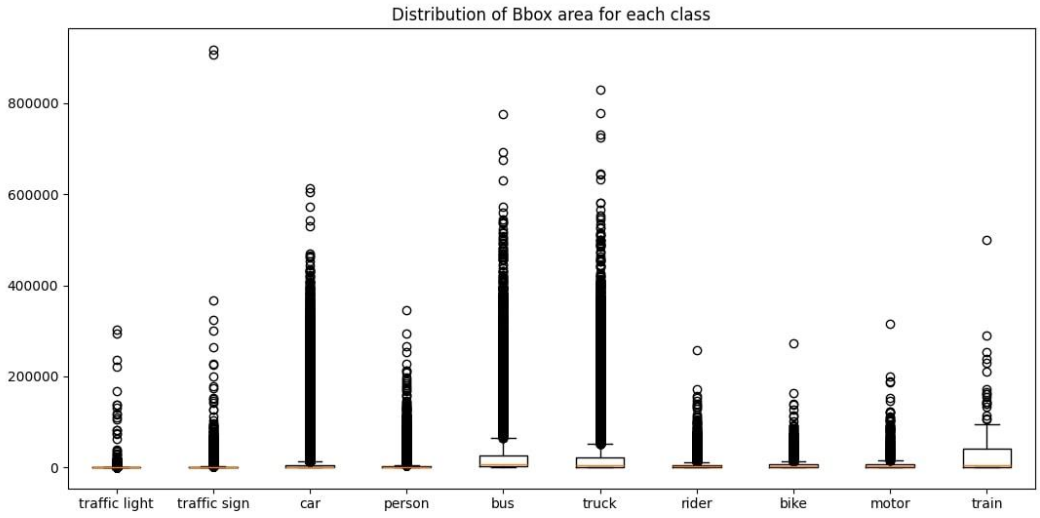
- The dataset is heavily skewed towards "not_truncated" instances, which might bias the model and hinder its performance on truncated objects.
- The presence of a significant number of truncated objects suggests that the model will need to be robust to handle partially occluded objects.
- To mitigate the class imbalance and improve the model's performance on truncated objects, data augmentation techniques like random cropping or occlusion can be applied.

6. **Instances of Occlusion annotations**



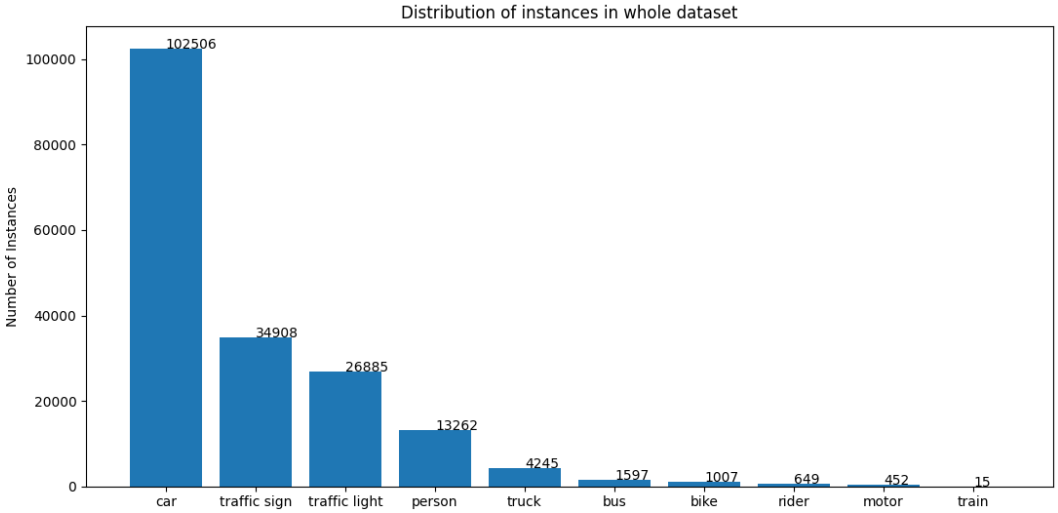
- The dataset shows almost class balance between "not_occluded" and "occluded" instances.
- The presence of a significant number of occluded objects suggests that the model will need to be robust to handle partially occluded objects. It will need to learn to identify objects even when parts of them are hidden by other objects or background elements.

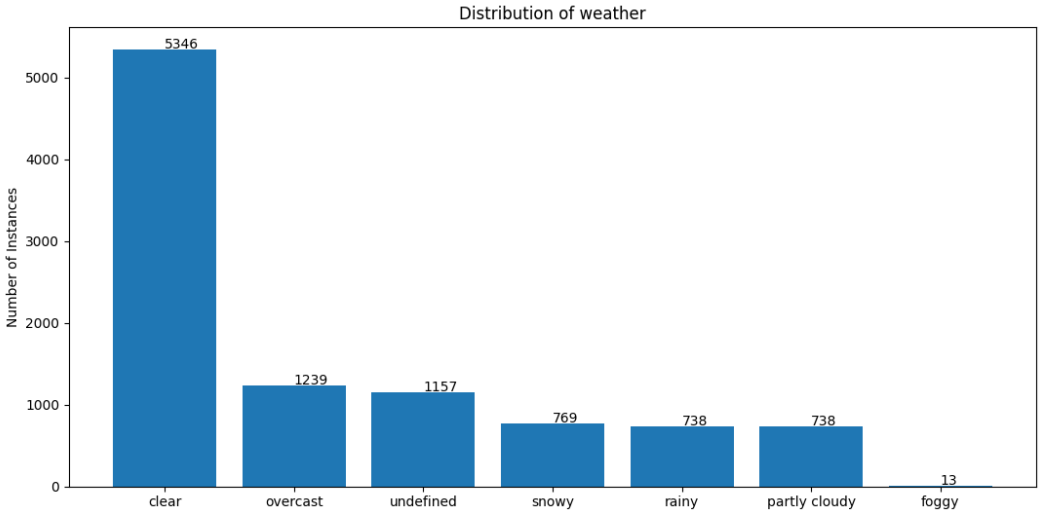
7. **Distribution of bounding box area for each class**

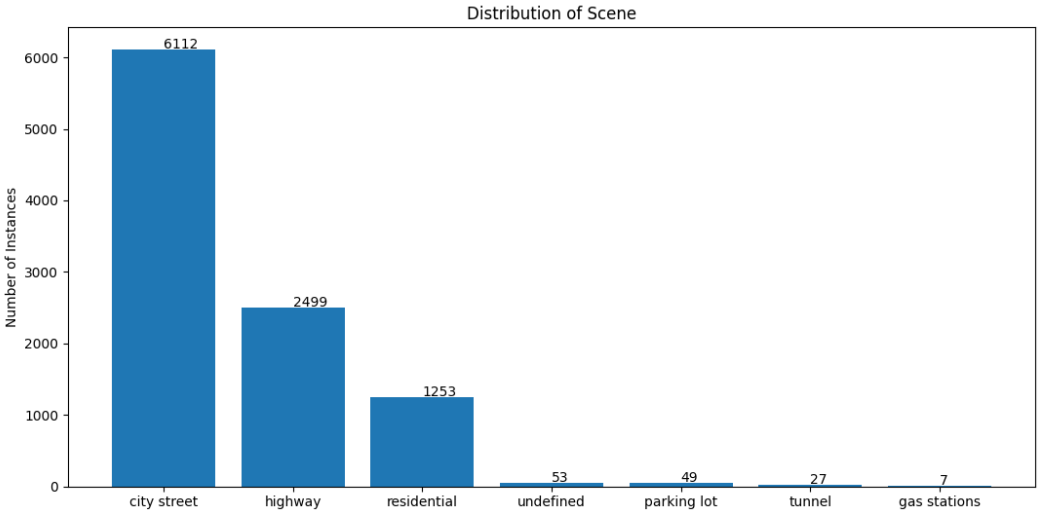


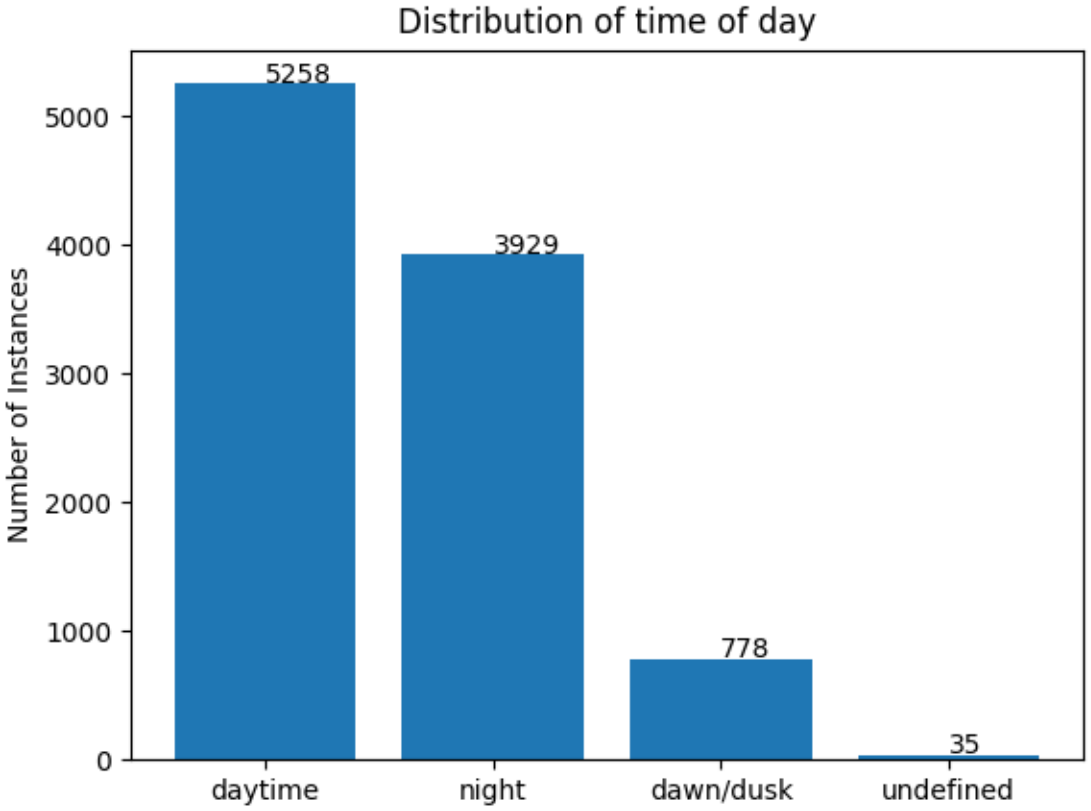
- The plot shows significant variation in the size of bounding boxes (BBoxes) for different object classes. For example, "traffic light" and "traffic sign" have much smaller BBoxes compared to "car" and "person". This highlights the need for a model that can effectively detect objects of varying sizes.
- The presence of outliers in some classes, such as "car" and "person," suggests that there might be some extremely large or small objects in the dataset. These outliers could potentially impact the model's training and performance.
- The distribution of BBox sizes within each class varies. Some classes, like "traffic light" and "traffic sign," have a more concentrated distribution, while others, like "car" and "person," have a wider range of sizes. This suggests that the model will need to be robust to handle different object sizes within each class.
- The presence of outliers and the variation in BBox sizes could indicate potential issues with data quality, such as incorrect annotations or inconsistent image resolutions. It's important to ensure data quality to train a reliable model.

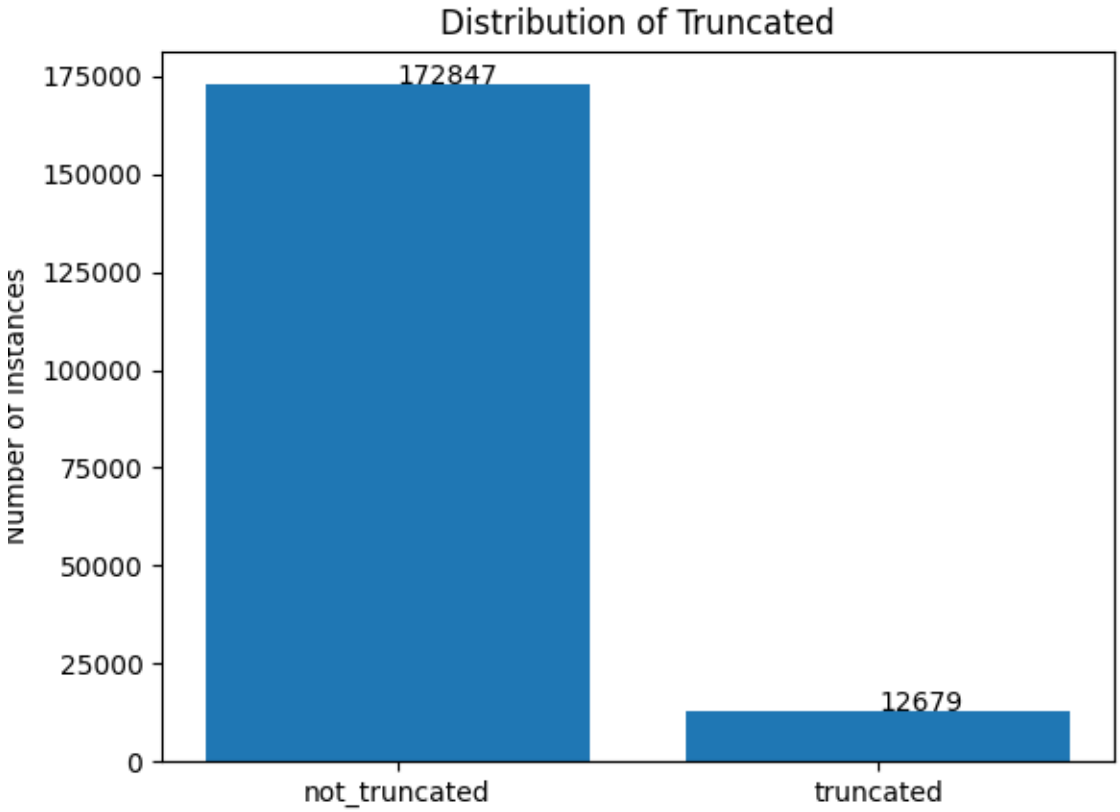
Analysis on Validation Set

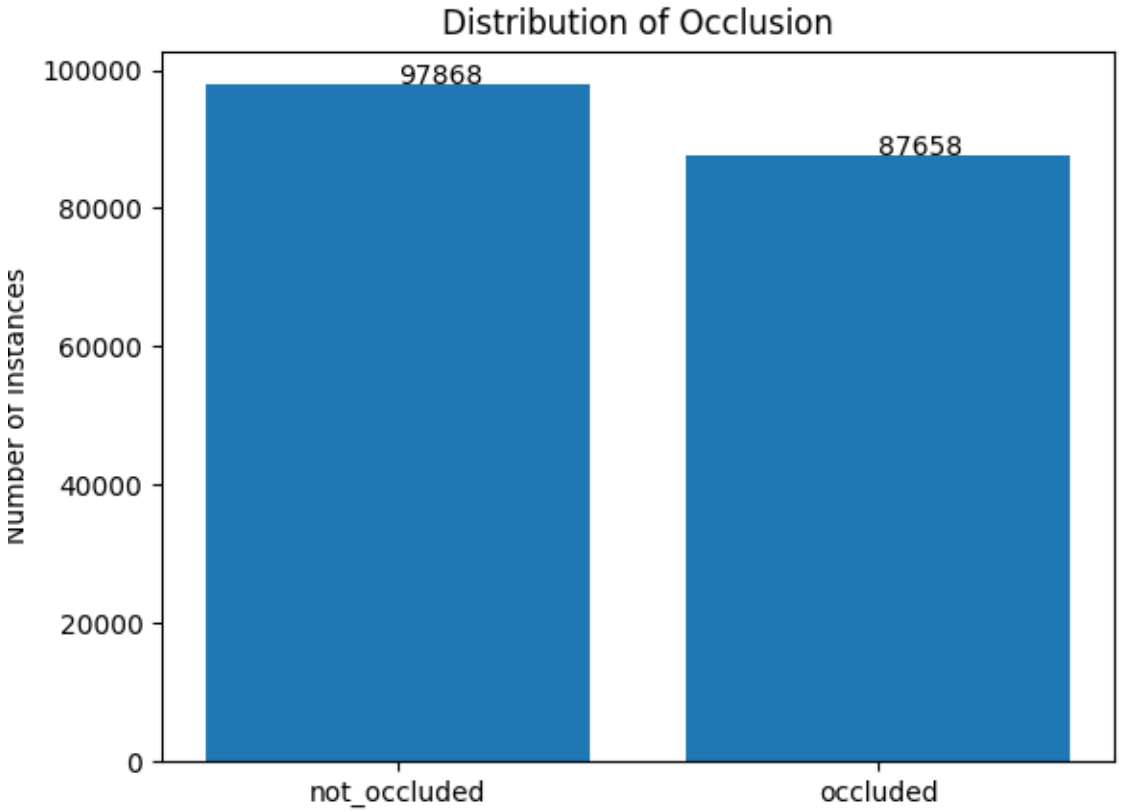


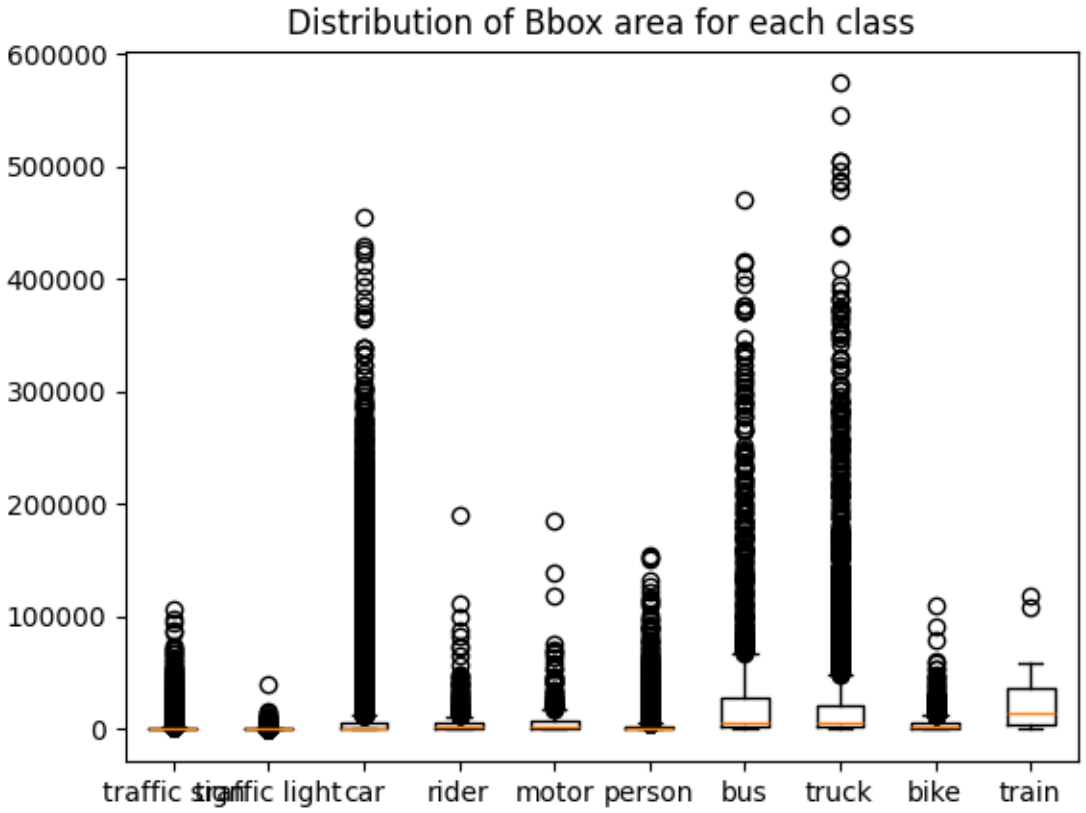



Distribution of Scene
Number of Instances
6112
2499
1253
53
49
27
7
city street
highway
residential
undefined
parking lot
tunnel
gas stations


Distribution of time of day
Number of Instances
5258
3929
778
35
daytime
night
dawn/dusk
undefined


Distribution of Truncated
Number of instances
175000
150000
125000
100000
75000
50000
25000
0
172847
12679
not_truncated
truncated


Distribution of Occlusion
Number of instances
100000
80000
60000
40000
20000
0
97868
87658
not_occluded
occluded


Distribution of Bbox area for each class
600000
500000
400000
300000
200000
100000
0
traffic sign
traffic light
car
rider
motor
person
bus
truck
bike
train