2025-2026

UTMIST

SceneClarity

Scene Reliability Scoring and Classification in Autonomous Vehicle (AV) Perception

Lead: Javier Huang

Developers: Ciny Liang, Tommy Yu, Ashwin Santhosh, Emmanuel Ko, Winston Liang, Elena Zhu, Isaac Wan, Gerardus Raynard Effrien

The Mission

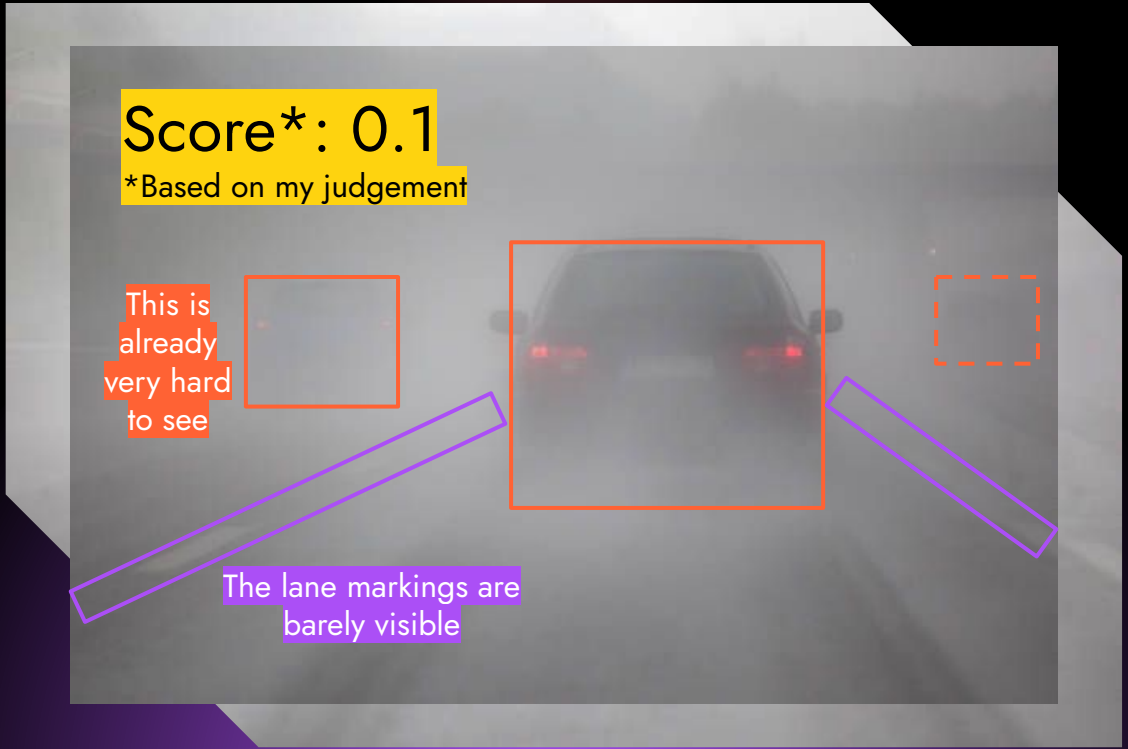
Autonomous vehicles depend on **computer vision** for real-time scene understanding, but performance often **degrades in adverse conditions** such as glare, fog, rain, or snow. This project aims to build a **modular reliability scoring pipeline** that quantifies and explains potential perception failures caused by environmental or visual degradation. Such a system will support decision-making, **safety assessment**, debugging, and driver notifications.

Part 1: Reliability Scoring

If you were driving, how **reliable** would you say this scene is?

Rate from 0 to 10.

In this project, our rating is from 0 to 1.



Part 2: Cause Attribution

What is the root cause of the issue?

A reason needs to be provided to justify the score.

We will focus on glare, weather, and time of day

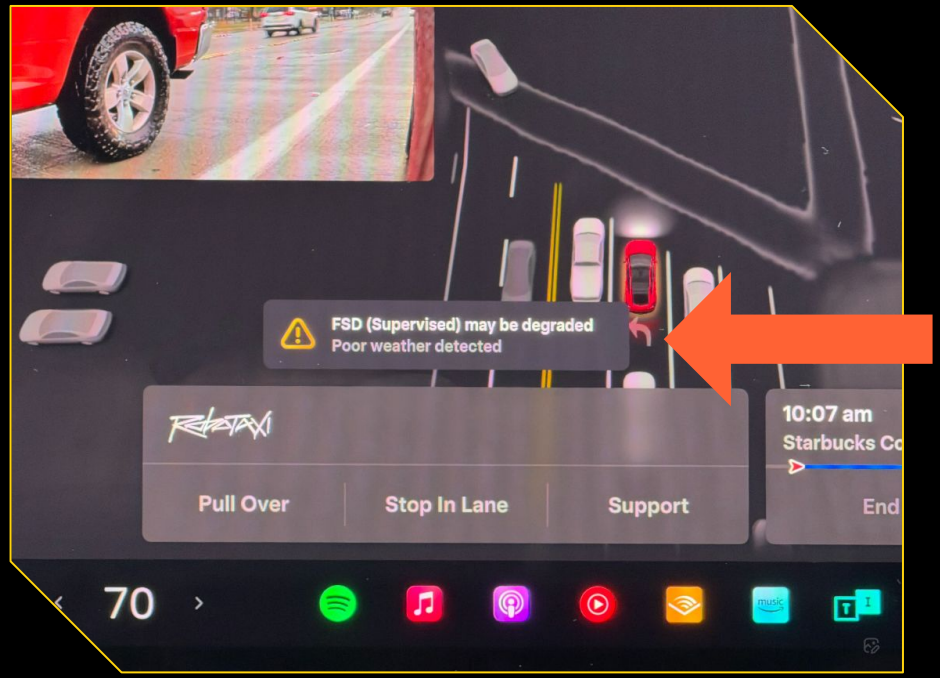
Intuitive for Humans

While the result felt intuitive, our estimates were unconsciously shaped by various tools.

These metrics are essential for AV decision-making and for diagnosing anomalies or failures.

1. Able to identify vehicles?
2. Able to identify lanes?
3. Road surface condition visible?
4. Able to read traffic signs?
5. Able to spot pedestrians and obstacles?

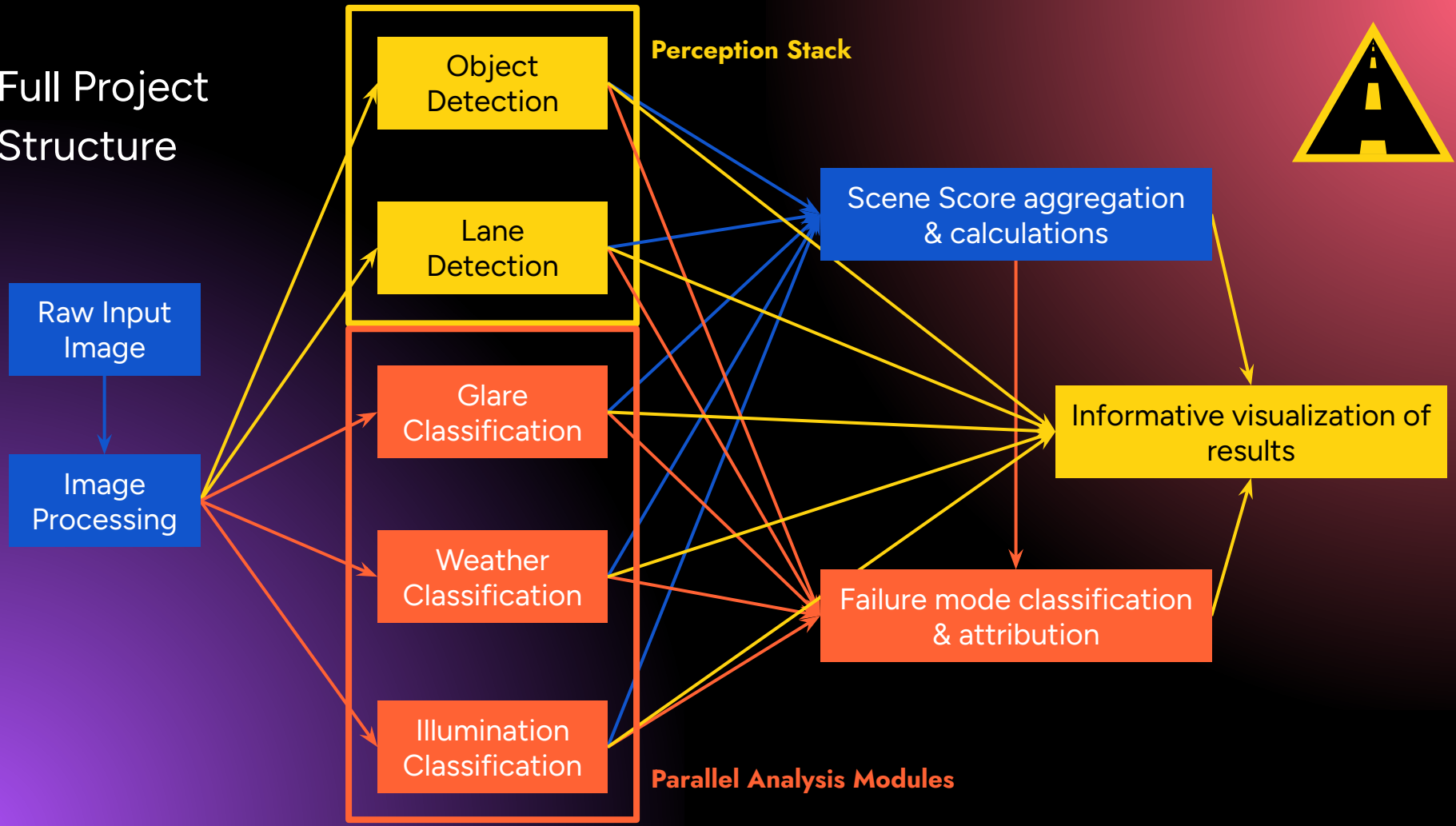
Real-world Example



A recent example (Jan 23, 2026) of Tesla's **Full Self Driving (FSD)** in its Robotaxi service with an error message stating the presence of poor weather. This, however, gives **limited interpretability**, as it is unclear what kind of weather it is. SceneClarity aims of combat this by providing the **exact condition** causing reliability degradation.

Image Source

Full Project Structure



Machine Learning Models

Perception Stack

- Object Detection
  - YOLOv8
  - YOLOv11
- Lane Detection
  - LaneNet

Classification

- Glare Classification
- Weather Classification
- Time-of-Day Classification

ResNet-18
ResNet-50
MobileNetV3
EfficientNetB2

Why does this matter?

Reliability and robustness

- Enables system-level reliability assessment instead of fragmented per-prediction confidence
- Captures interactions between multiple failure sources in complex scenes
- Allows early detection of unreliable perception under distribution shift

Interpretability and analysis

- Improves interpretability via factorized attribution of degradation causes
- Facilitates root-cause analysis and faster debugging of failures
- Bridges the gap between uncertainty estimation and actionable monitoring

Safety and decision-making

- Supports safety-critical decisions about when to trust or override perception outputs

Integration and scalability

- Provides a drop-in diagnostic layer without modifying existing models
- Scales across different sensors and model architectures through modular design

BDD100K

Our work with the BDD100K Dataset

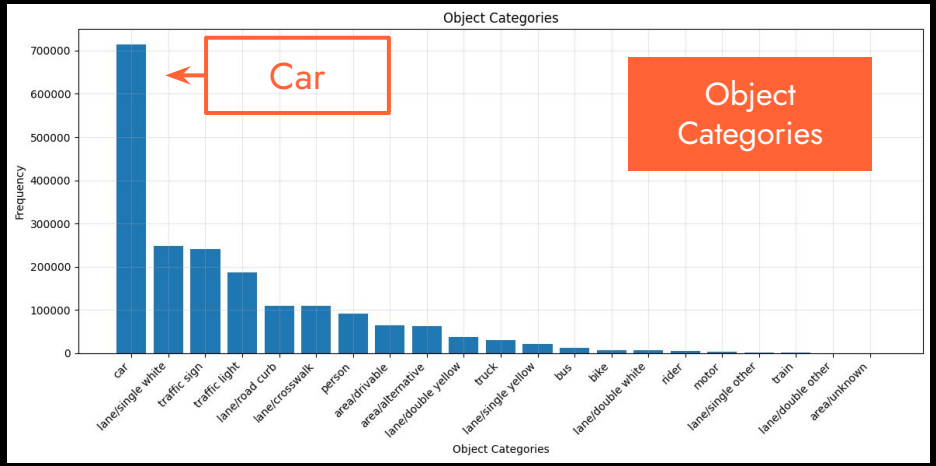
BDD100K Data Images

BAIR

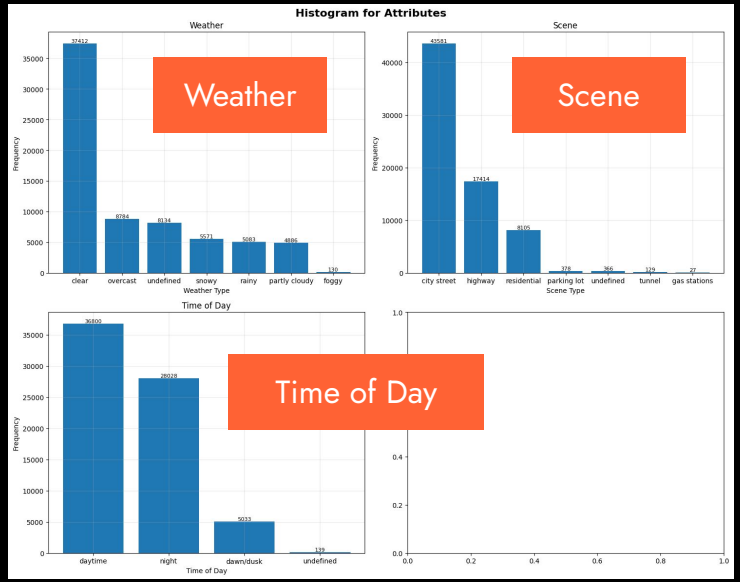
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

BDD100K Dataset EDA

On average, 27.8 objects in 21 categories are detected from each of the 70000 images.



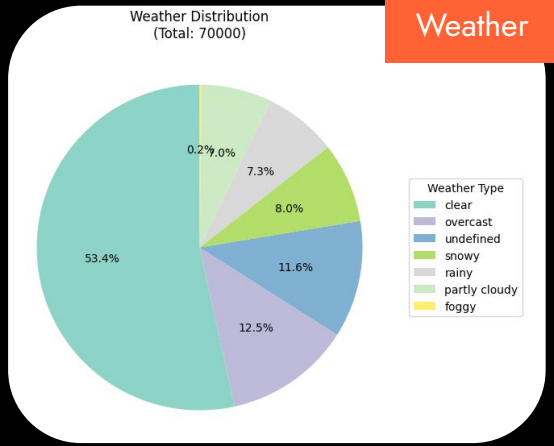
Other attributes of the images, including weather, scene, and time, are also recorded.



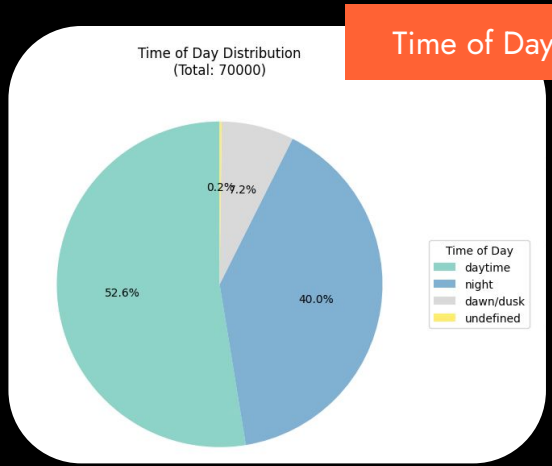
BDD100K Dataset EDA

Distributions

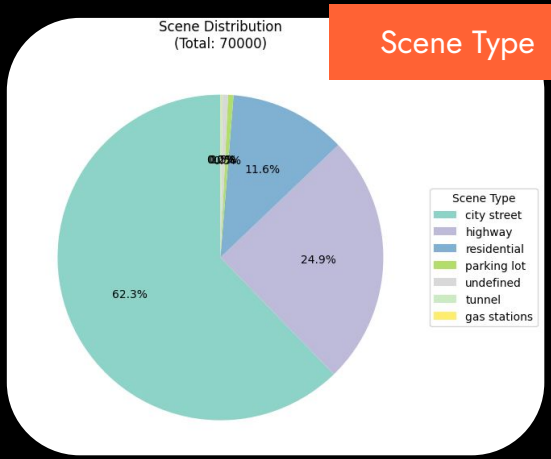
Weather



Time of Day

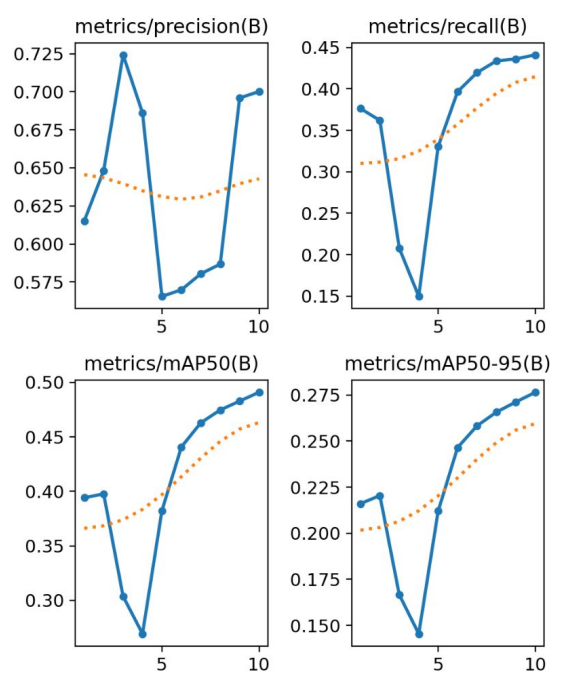


Scene Type



YOLOv8 Results

Tiled Training Images (640x640)

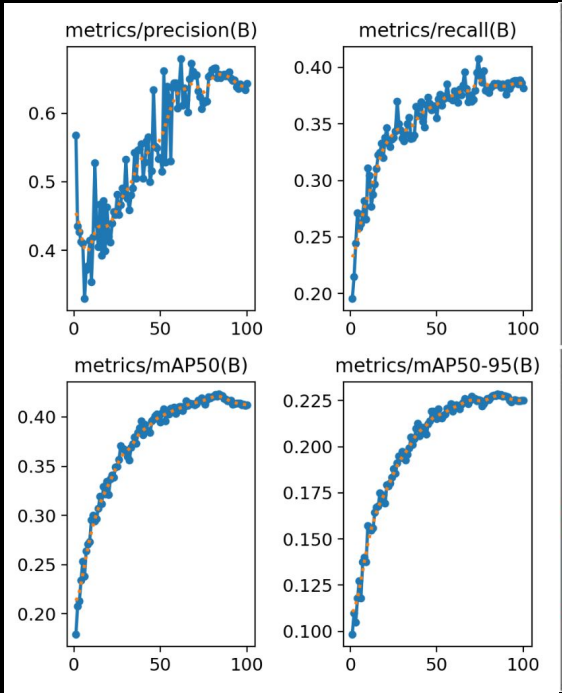


Metrics

Validation Prediction Examples

YOLOv8 Results

Non-tiled Training Images (YOLO internal resizing)

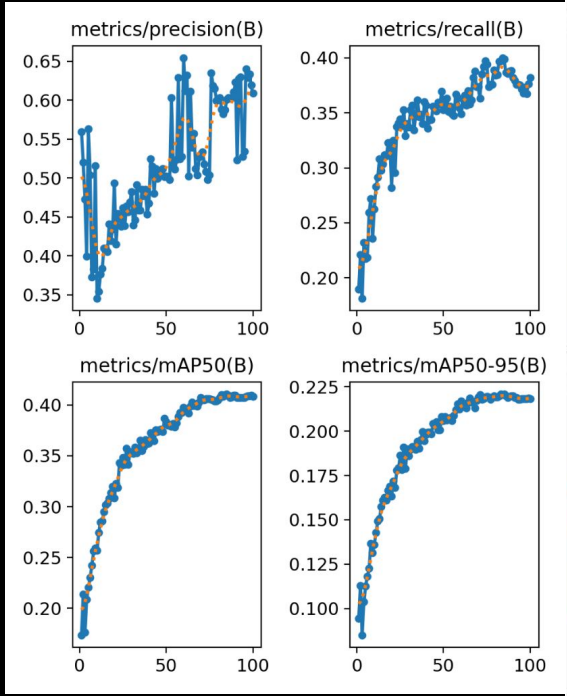


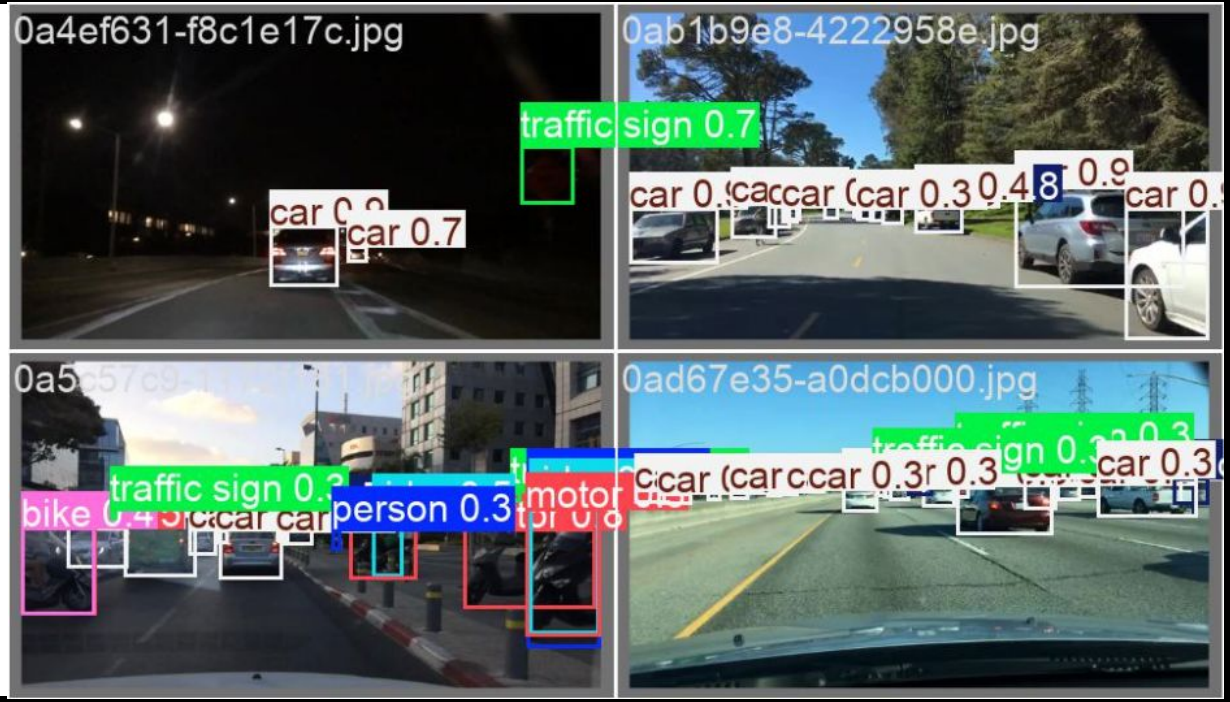
Metrics

Validation Prediction Examples

YOLOv11 Results

Non-tiled Training Images (YOLO internal resizing)





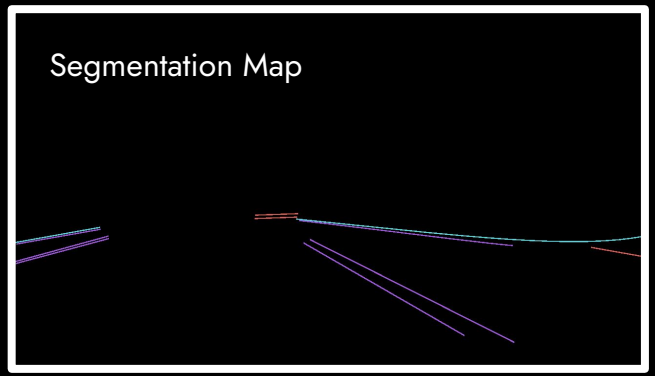
Metrics

Validation Prediction Examples

Lane Segmentation

EDA + Format Conversion

+



Glare

Our work with glare augmentations

Initial Glare Augmentation Approach

Overlaying cropped glare images on brightest regions

Raw Image

Glare Overlay

Data Fidelity Issues

- The glare overlay color and bloom may be inconsistent with actual glare properties/geometry that would occur in reality.
- No scalable or simple way to do dynamic sizing of glare profiles since cropped glare images correspond to specific glare types/conditions.
    - Glare as seen by camera include Lens Flare, Veiling Glare, Specular Reflection Glare, and Sensor Blooming to name a few.
    - eg. Would want to avoid putting specular reflection glare on a neon sign (bloom is more appropriate). *Observe cases below!*

Specular Reflection from Car Paint

Bloom from Neon Sign

Glare Augmentation (Version 1)

Glare simulation with three bloom levels

- Applied the artificial glare method to 500 dataset images
- 3 levels of glare, determined by 2 parameters:
  - `k_bloom`
    - Blur radius size → wider or tighter glow
  - `bloom_intensity`
    - Blend strength → brighter or subtler glow

Acceptable Results

Glare Augmentation (Version 2)

1. Bright spot locator
2. Streaks on bright spot
3. Additional “reflection” glare in a random spot
4. Lens flare placed radially opposite to a bright spot

Realistic Results!

Glare Dataset Packaging

+

Label = 0

Label = 1

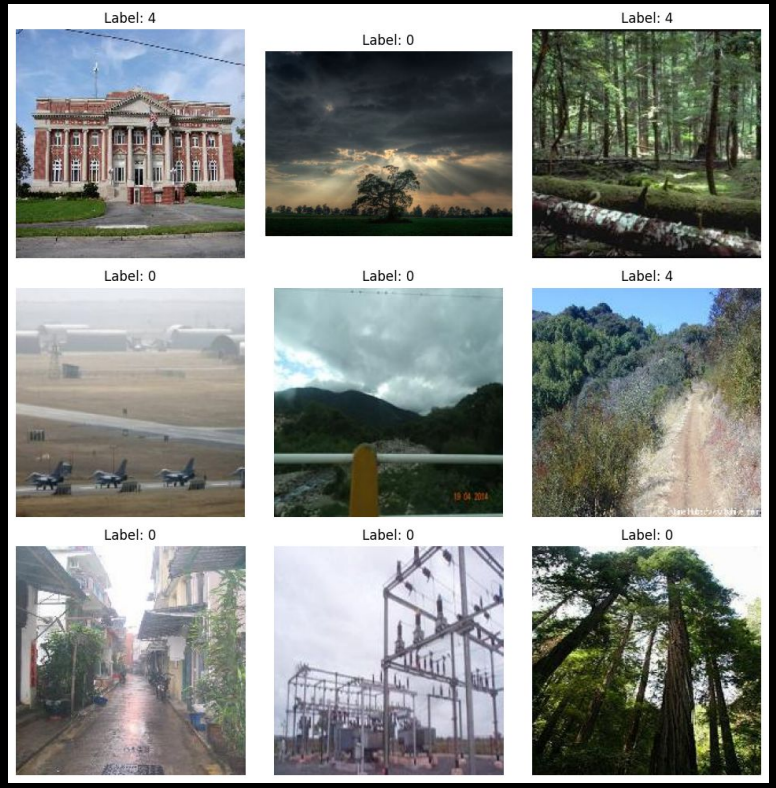
Glare Augmented Dataset

Images: *Binary Classification* - *_no_glare.jpg (label 0) and *_glare.jpg (label 1)
For each original image, create a paired glare version → doubles the dataset size
Labels: labels.json (full metadata, including bloom parameters) and labels.csv (image, label, class)

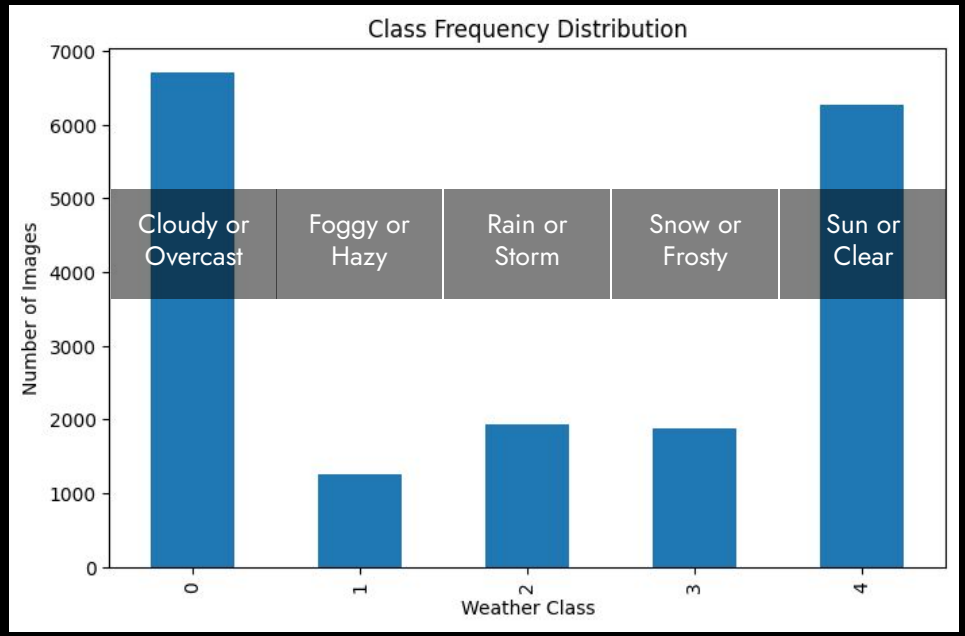
WeatherNet

Our work with the WeatherNet Dataset

WeatherNet



WeatherNet-05 is a weather image classification dataset consisting of **18,039** images labeled into **5 distinct weather-related classes**. The dataset is suitable for training and evaluating computer vision models on the task of classifying **weather conditions** based on image data.



**Margin mean is the average distance between the chosen prediction and its closest alternative.*

ResNet-18 vs ResNet-50

Both trained under the **WeatherNet** Dataset

`num_epochs=10, lr=0.001, batch_size=32`

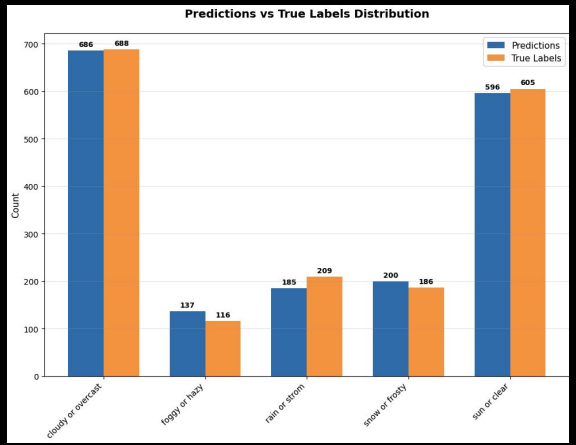
ResNet-18

Accuracy: 88.11%

Macro F1: 0.870

Confidence Mean: 0.94

Margin Mean*: 0.89



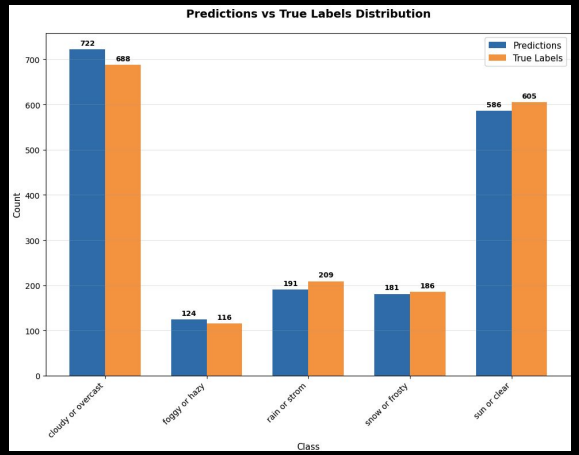
ResNet-50

Accuracy: 89.48%

Macro F1: 0.897

Confidence Mean: 0.95

Margin Mean*: 0.91



Predictions
True Labels

Example from ResNet-50

**Margin mean is the average distance between the chosen prediction and its closest alternative.*

EfficientNet B1

Trained under the **WeatherNet** Dataset

`num_epochs=10, lr=0.001, batch_size=32`

ResNet-50

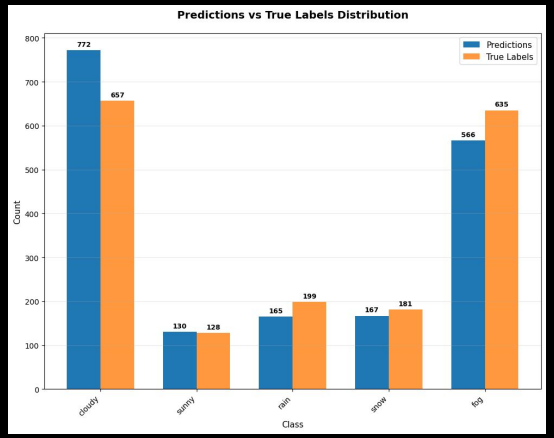
Accuracy: 89.48%
Macro F1: 0.897
Confidence Mean: 0.95
Margin Mean*: 0.91

ResNet-18

Accuracy: 88.11%
Macro F1: 0.870
Confidence Mean: 0.94
Margin Mean*: 0.89

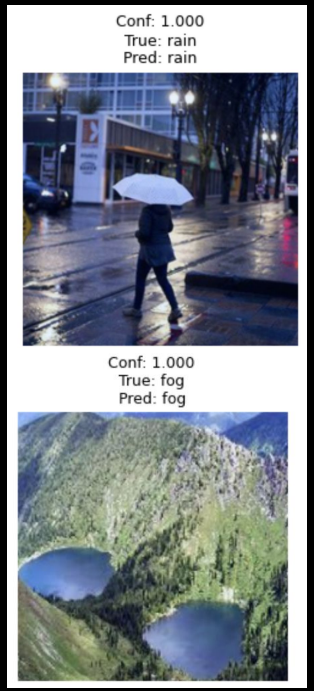
EfficientNet B1

Accuracy: 79.21%
Macro F1: 0.804
Confidence Mean: 0.92
Margin Mean*: 0.85



Predictions
True Labels

Example from EfficientNet B1



Dark Zurich

Our work with the Dark Zurich Dataset

Dark Zurich

The *available dataset involves the day and night images taken from 151 different locations. It has equivalent amounts of daytime and nighttime images (151 each)

Path name: GP010364_frame_000009_rgb_anon.png
Label: night

Path name: GP010362_frame_000554_rgb_anon.png
Label: day

**The dataset available on Kaggle only contains 151 day and nighttime images. The original dataset is not accessible*

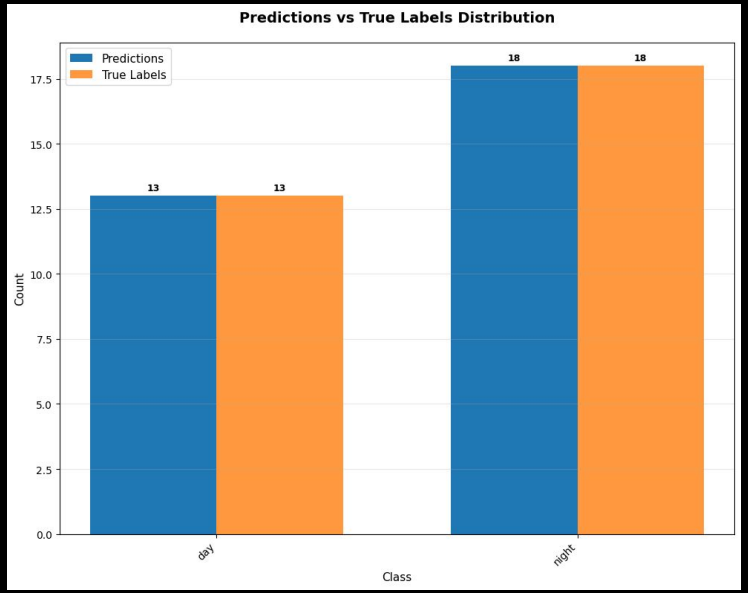
ResNet-18, ResNet-50, MobileNet, EfficientNetB1 Collective Results

Both trained under the **Dark Zurich** Dataset

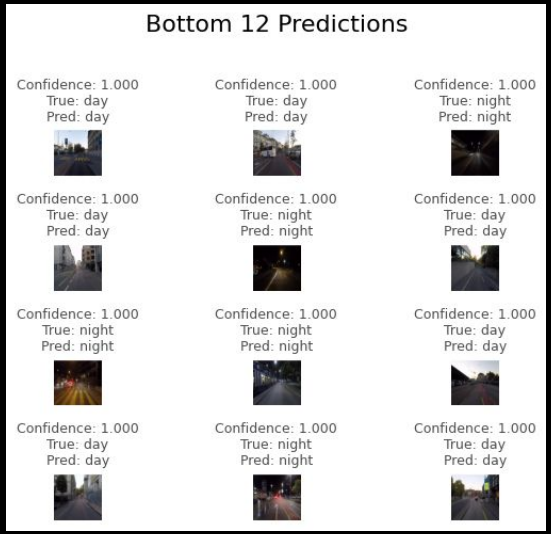
`num_epochs=10, lr=0.001, batch_size=32`

All models

- Scored 100% accuracy
- Low margin and entropy with high confidence



Example from ResNet-18

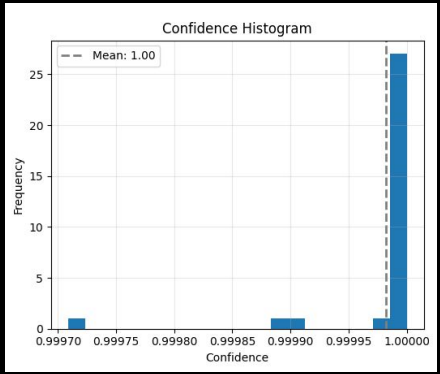


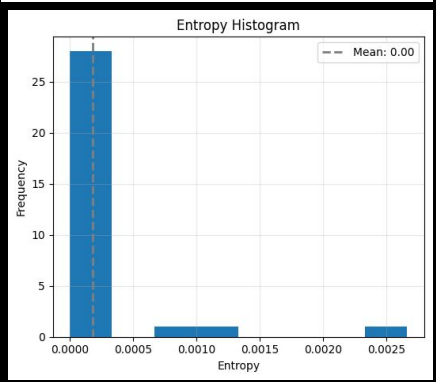
Dark Zurich

Detailed Results of Models

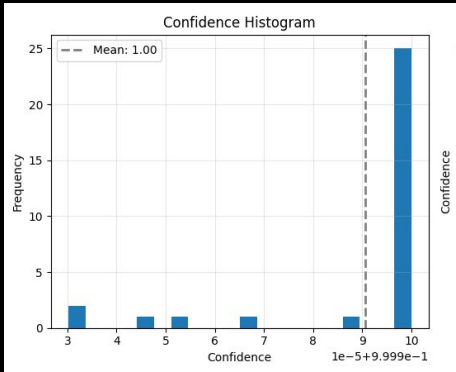
All trained under the **Dark Zurich** Dataset
Confidence and Entropy Histograms

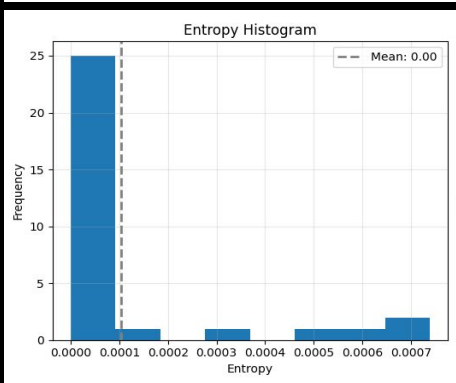
ResNet-18



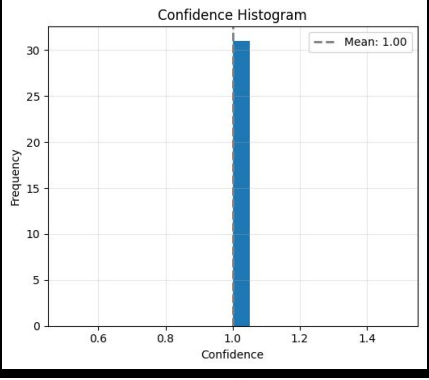


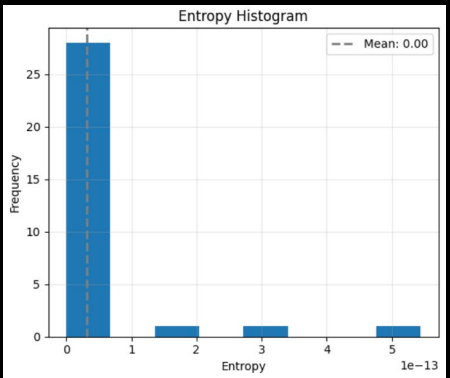
ResNet-50





MobileNet





EfficientNetB1