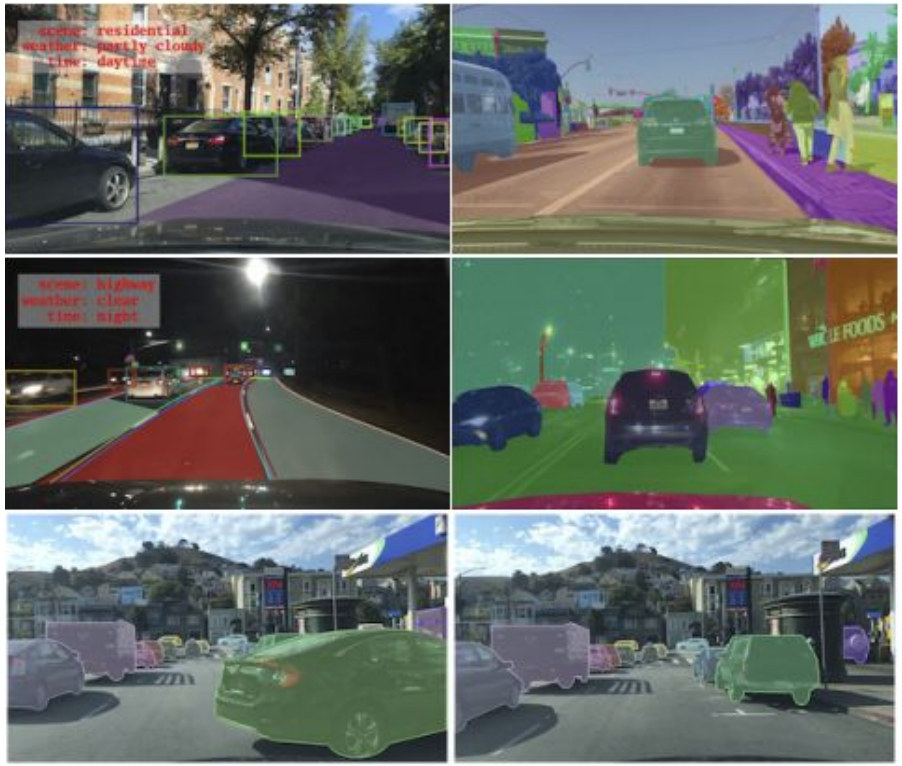
BDD100K - (Berkeley Deep Drive) dataset

Alexander Schad, Jonathan Jacobs

Dataset Background

- BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning - https://arxiv.org/pdf/1805.04687v2
- 100K driving videos (40 seconds each) collected from more than 50K rides, covering New York, San Francisco Bay Area
- Six weather conditions (Rain, Snow ...), six scene types, and three distinct times of day (Day, Night,Dawn/Dusk)

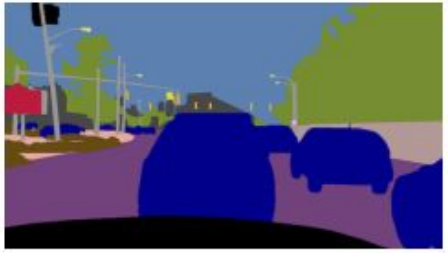


Computer Vision Tasks

- Image Tagging
- Lane Detection
- Drivable Area Segmentation
- Road Object Detection
- Semantic Segmentation
- Instance Segmentation
- Multi-object Detection Tracking
- Multi-object Segmentation Tracking
- Domain Adaptation
- Imitation Learning

Lane Marking Annotation

Drivable Area Annotation

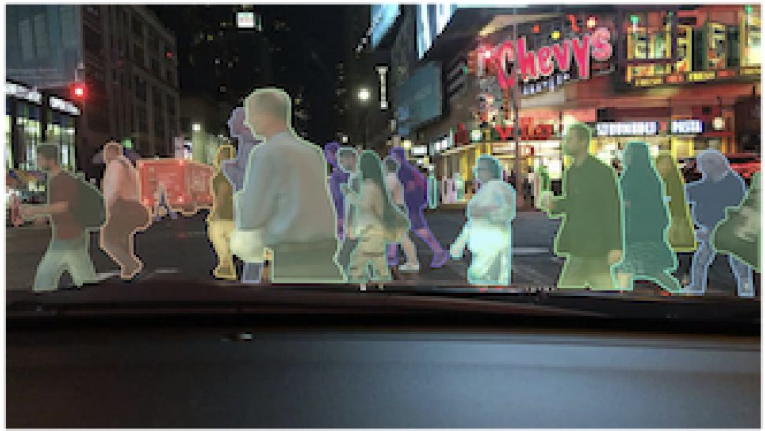


Semantic Segmentation

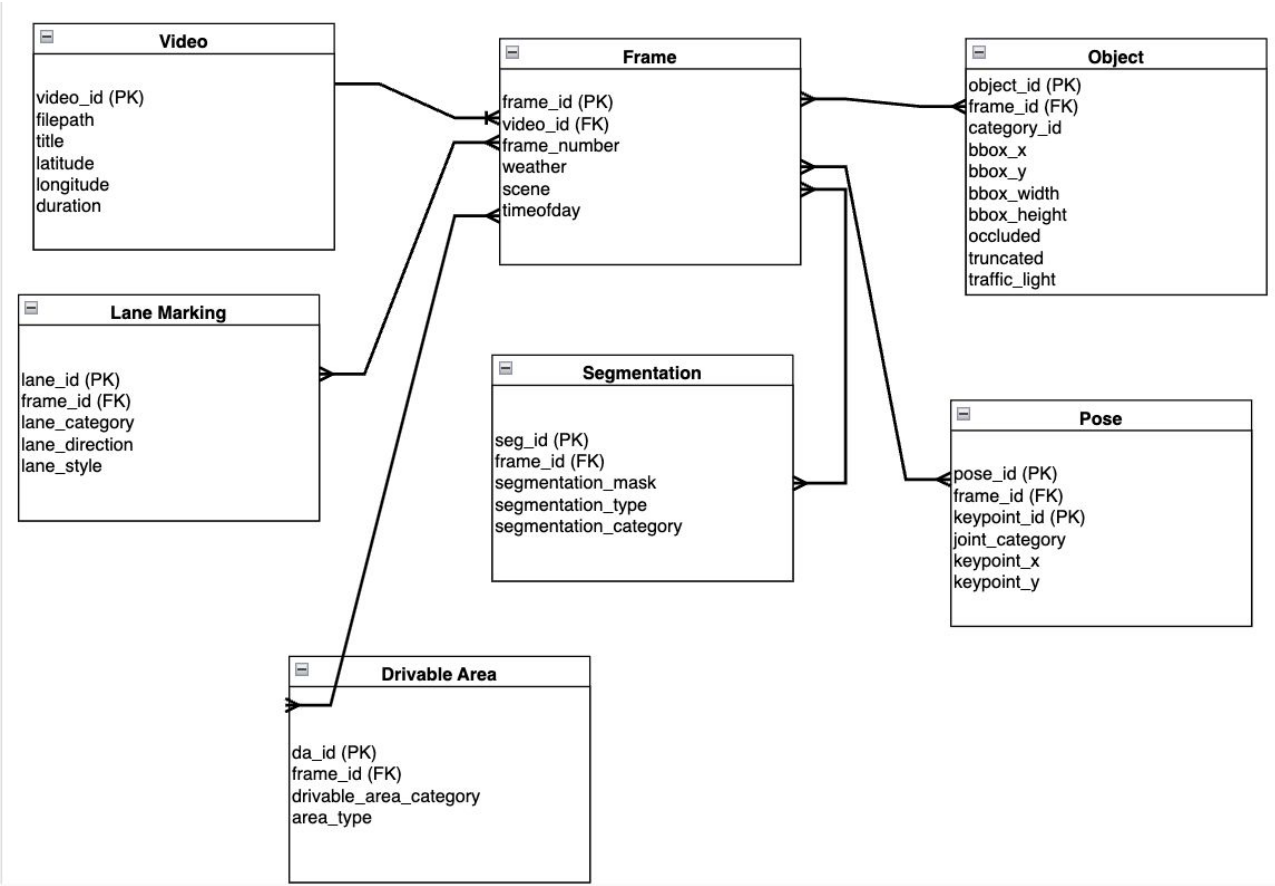
Object Detection

Data Governance Issues

- The BDD100K dataset contains various types of sensitive information, including GPS data, video footage, and personal identifiers. These data points could potentially be used to identify individuals or specific locations, posing significant privacy risks.
- To mitigate these risks, several protection measures should be implemented:
    1. **Anonymization**: Personal identifiers and location details should be anonymized to prevent identification.
    2. **Secure Storage**: Data should be stored in secure, encrypted databases with restricted access to authorized personnel only.



ER Diagram


Video
video_id (PK)
filepath
title
latitude
longitude
duration
Frame
frame_id (PK)
video_id (FK)
frame_number
weather
scene
timeofday
Object
object_id (PK)
frame_id (FK)
category_id
bbox_x
bbox_y
bbox_width
bbox_height
occluded
truncated
traffic_light
Lane Marking
lane_id (PK)
frame_id (FK)
lane_category
lane_direction
lane_style
Segmentation
seg_id (PK)
frame_id (FK)
segmentation_mask
segmentation_type
segmentation_category
Pose
pose_id (PK)
frame_id (FK)
keypoint_id (PK)
joint_category
keypoint_x
keypoint_y
Drivable Area
da_id (PK)
frame_id (FK)
drivable_area_category
area_type

Use Cases

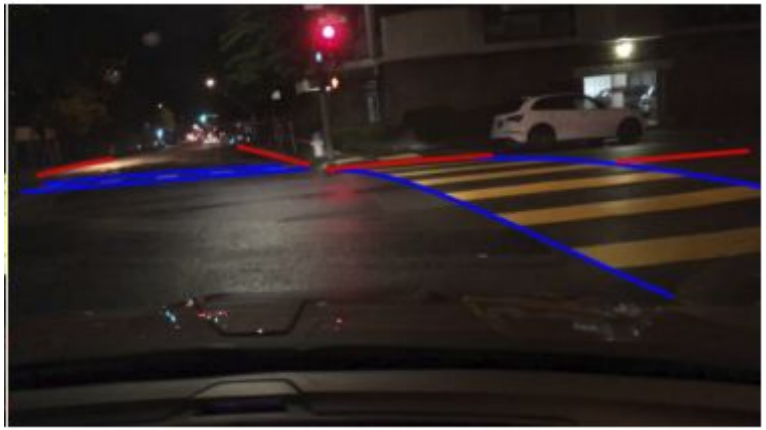
1. Counting the Number of objects for each frame in a given video (Frequently for a visualization application)

```sql
SELECT f.frame_number, COUNT(o.object_id) AS object_count
FROM Frame f
LEFT JOIN Object o ON f.frame_id = o.frame_id
LEFT JOIN Video v ON f.video_id = v.video_id
WHERE v.title = "NYC June 1 2019"
GROUP BY f.frame_number;
```

Use Cases

2. Find the number of frames with a crosswalk per video (Infrequently - Ad hoc)

```
SELECT v.video_id, count(l.lane_id)
FROM Frame f
LEFT JOIN Video v ON f.video_id = v.video_id
LEFT JOIN LaneMarking l ON f.frame_id = l.frame_id
WHERE l.lane_category = 1
GROUP BY v.video_id;
```



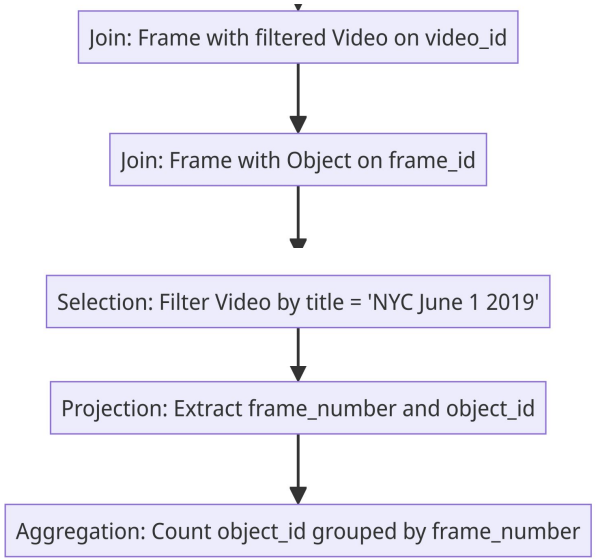
Use Cases

3. Finding the number of frames that are available for every time of day and weather combination (Infrequently - Ad hoc)

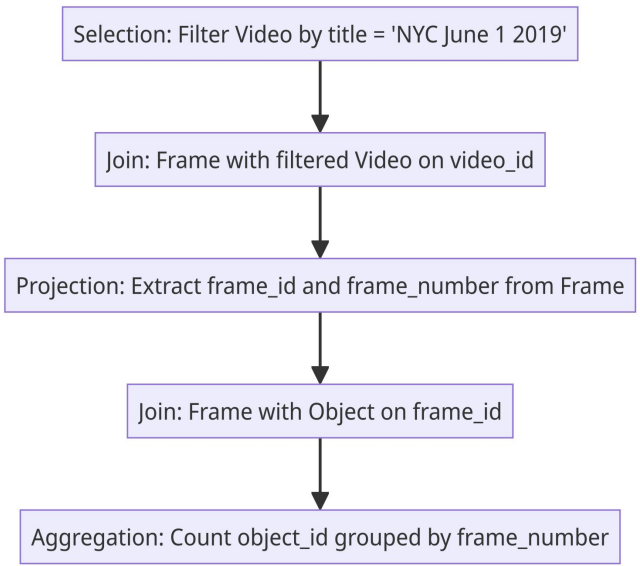
```
SELECT f.weather, f.timeofday, count(*)
FROM Frame f
LEFT JOIN Video v ON f.video_id = v.video_id
GROUP BY f.weather, f.timeofday;
```

RA Trees / Indexes

Naive



Optimized



Indexes:

- **Weather and Time of Day Index:** Unclustered Index on `weather` and `timeofday` columns in the `Frame` table to support the third query.
- **Frame ID Index**: Unclustered Index on `frame_id` column in the `Object` and `LaneMarking` tables to support the first and second queries.

DBMS Choice

PostgreSQL -

- The ER diagram above lends itself well to a relational databases as the original nested JSON format of the data made it hard to traverse down to specific annotation types
- Task specific annotations are easier to access via their own tables
- PostgreSQL's scalability features, such as parallel query execution and partitioning, make it suitable for managing the extensive BDD100K dataset.
- Additionally, its support for JSON data provides flexibility for future extensions.
- The strong community and range of extensions further enhance its functionality, making PostgreSQL an ideal choice for this project.

Questions?