Identifying and Analyzing Data Quality Issues in Autonomous Driving Datasets

Group Members – Group-6:

1. Aiswarya Goriparthi
2. Nitya Vattam
3. Vinuthna Reddy Vintha
4. Shanmukha Varma Kothapalli
5. Bhavya Madhuri Devarakonda
6. Pavani Kommineni

Course: INFO 5731 - Computational Methods for Information Systems

Date: December 1st, 2024

Table of Contents

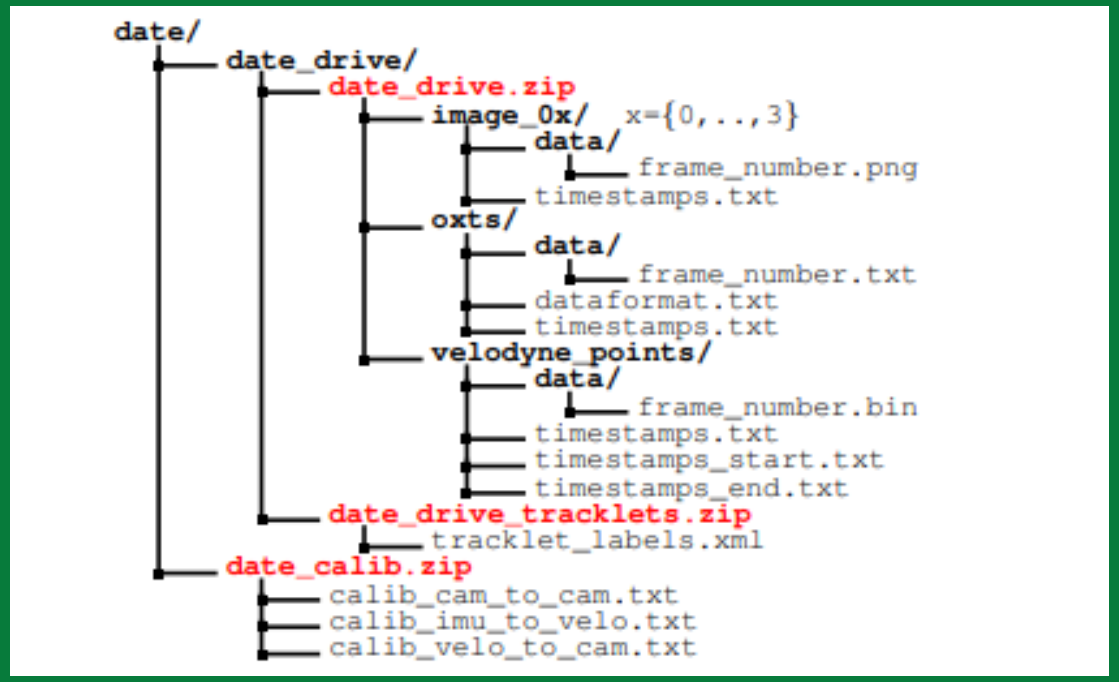
Background and Motivation

- Autonomous driving datasets are important in training as well as testing AI models.
- Poor quality of data leads to serious issues in model performance and user safety.
- To systematically assess the KITTI dataset for the data quality issues and suggest actionable insights

Research purpose

- Detect and analyze data quality issues in the KITTI dataset.
- Assess datasets based on:
    - Completeness
    - Consistency
    - Correctness
    - Redundancy
- Improve the reliability of AI models in autonomous driving.
- Instruct better preparation and usage practices of datasets.

Dataset structure



Dataset structure

1.Image Data : Have rectified grayscale images in the PNG format for object detection as well as for tracking tasks.
2.Velodyne Point Cloud Data : Got 3D spatial data along with coordinates and reflectance values, also critical for 3D perception.
3.Calibration Files : Have sensor parameters of calibration.
4.Annotations and Timestamps: Got annotations of 3D object

Methodology

- **Completeness:**
  - Check for missing data: timestamps, tracklets, images.
  - Identify temporal gaps in timestamps.
- **Consistency:**
  - Assess relationships between datasets, like timestamps versus tracklets.
  - Review tracking levels and alignment across modalities.
- **Correctness:**
  - Find negative dimensions of objects' bounding boxes.
  - Identify invalid class IDs or other missing data fields.
- **Redundancy:**
  - Calculate file hashes and look for identical files.

Tools and Techniques

- **Tools Used**:
- **Programming**: Python, Pandas, Matplotlib, NumPy.
- **Data Handling**: parsing of XML, processing of csv, hashing - file redundancy.
- **Visualization**: Seaborn for distribution of data and for timestamp gaps.
- **Automation**: Developed reusable scripts to check data quality.

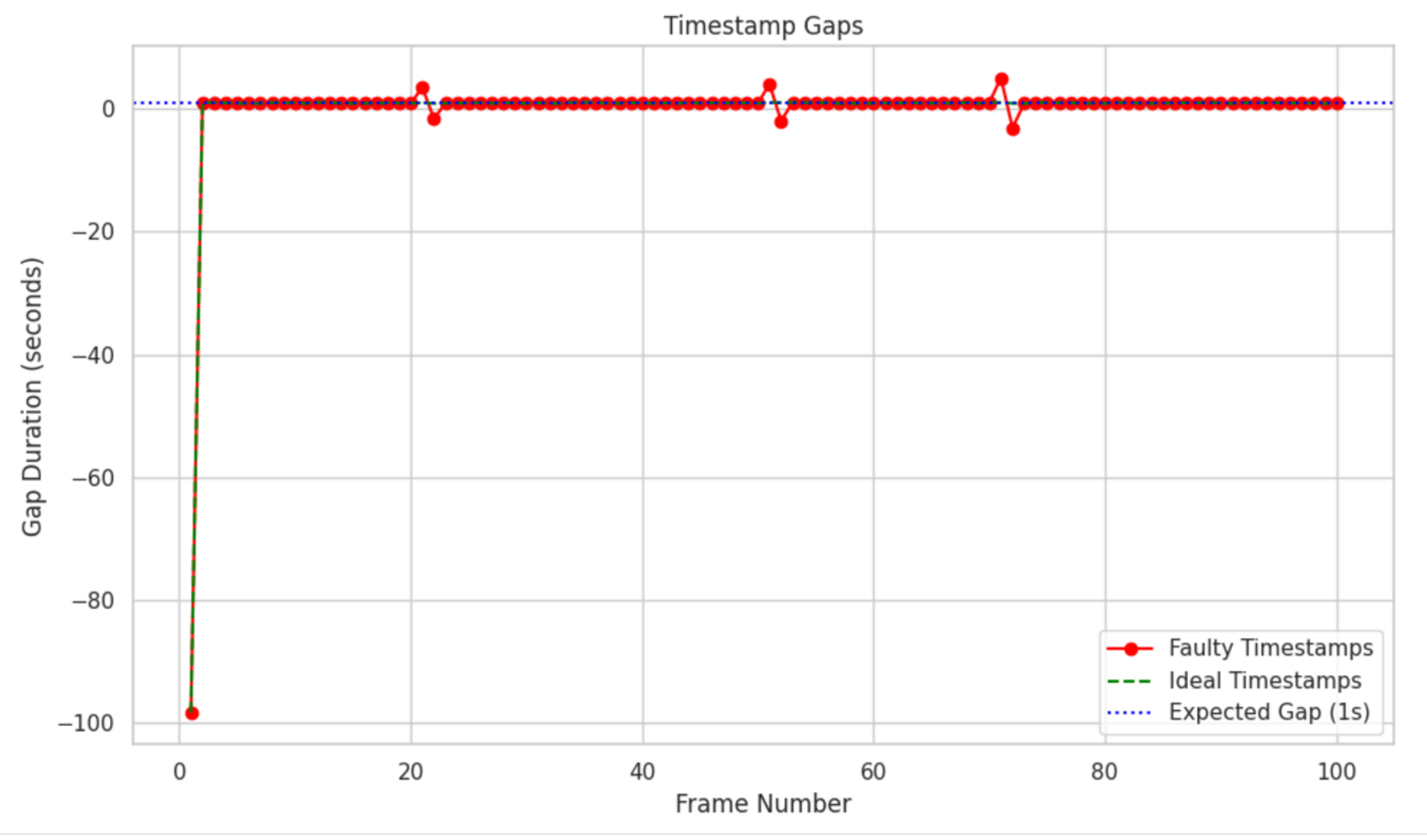
Key Findings

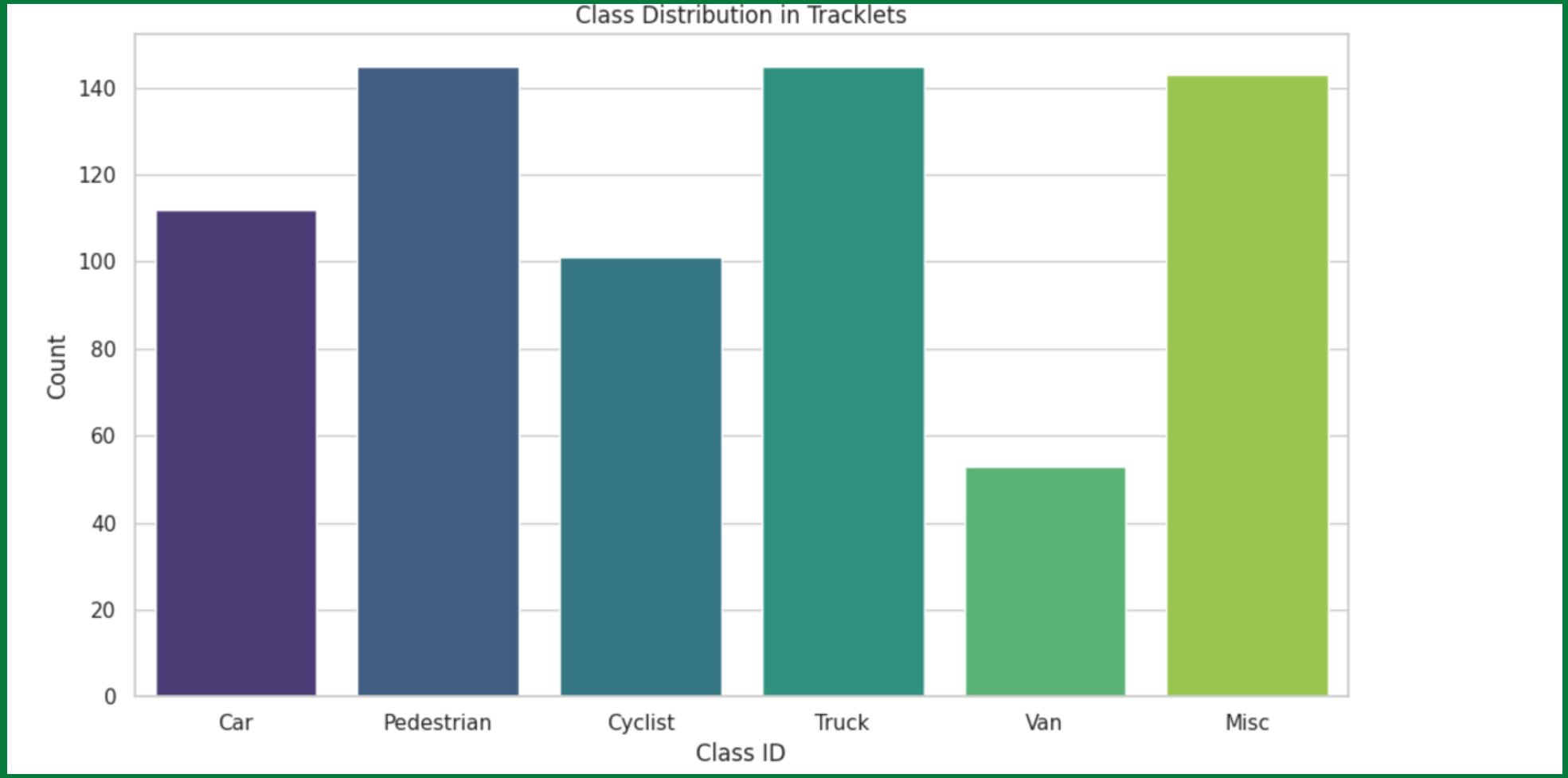
- **Completeness:**
  - Temporal Gaps: More than 2 seconds in a few sequences.
- **Consistency:**
  - Invalid tracking levels: Nonstandard values detected
- **Correctness:**
  - Object bounding boxes containing negative dimensions
  - Invalid class IDs
- **Redundancy:**
  - Multiple instances of same image files present

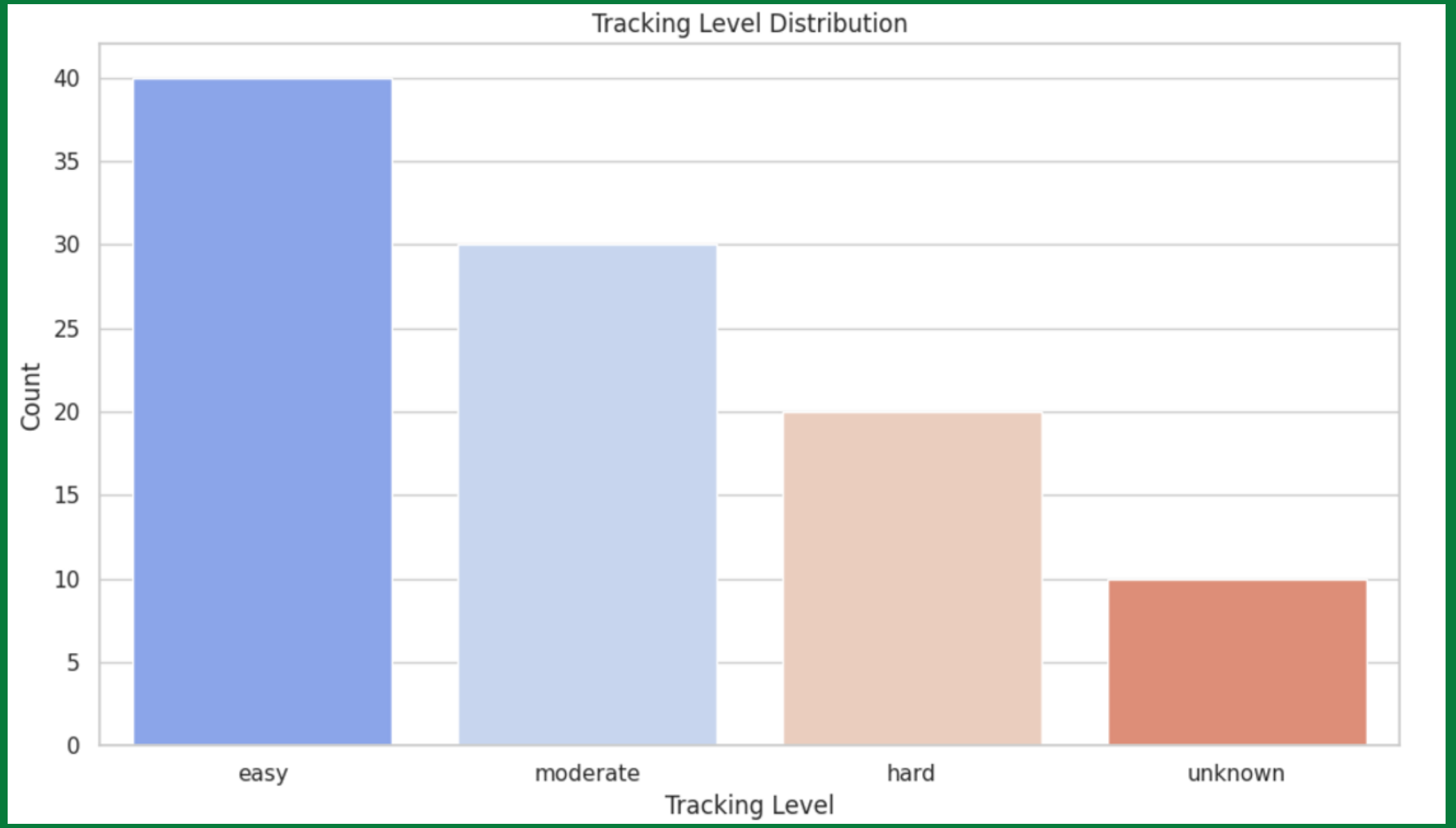
Results

```
Total inconsistent frames: 56

Processing Residential Dataset...

Downloading 2011_09_26_drive_0019_sync.zip for Residential dataset...
Downloaded dataset.
Unzipping Residential dataset...
Extracted dataset.
Located `image_02` data folder for Residential at: ./kitti_data/Residential/2011_09_26/2011_09_26_drive_0019_sync/image_02/data
Expected frame count from timestamps file: 481

Running Redundancy Check for Residential (Week 2)...
Total duplicates found: 0

Running Completeness Check for Residential (Week 3)...
Expected frames: 481, Actual frames: 481
Missing frames: 0

Running Temporal Consistency Check for Residential (Week 4)...
Inconsistency detected between frame 0 and frame 1 with diff 116.93476757917338
Inconsistency detected between frame 1 and frame 2 with diff 117.01004544641259
Inconsistency detected between frame 6 and frame 7 with diff 119.17236285560924
Inconsistency detected between frame 7 and frame 8 with diff 126.94013097155126
Inconsistency detected between frame 8 and frame 9 with diff 126.5016969046341
Inconsistency detected between frame 9 and frame 10 with diff 122.76783896940418
Inconsistency detected between frame 13 and frame 14 with diff 117.81350867775988
Inconsistency detected between frame 14 and frame 15 with diff 122.54038504204688
Inconsistency detected between frame 15 and frame 16 with diff 122.57432170334586
Inconsistency detected between frame 16 and frame 17 with diff 122.4609647521918
Inconsistency detected between frame 22 and frame 23 with diff 118.30663088208982
```

Results


Timestamp Gaps
Gap Duration (seconds)
0
−20
−40
−60
−80
−100
0
20
40
60
80
100
Frame Number
Faulty Timestamps
Ideal Timestamps
Expected Gap (1s)


Class Distribution in Tracklets
Count
0
20
40
60
80
100
120
140
Car
Pedestrian
Cyclist
Truck
Van
Misc
Class ID


Tracking Level Distribution
Count
40
35
30
25
20
15
10
5
0
easy
moderate
hard
unknown
Tracking Level

Redundancy Table:

	Original File	Duplicate File
0	000001.png	000001_copy.png
1	000015.png	000015_copy.png

Correctness Issues:

	id	error
0	1	Negative dimensions (length=-1.5)
1	5	Invalid class_id (-1)

Timestamp Gaps Table:

	Frame	Gap Duration (s)	Type
0	20	2.5	Temporal Gap
1	50	3.0	Temporal Gap
2	70	4.0	Temporal Gap

Summary of Data Quality Issues:

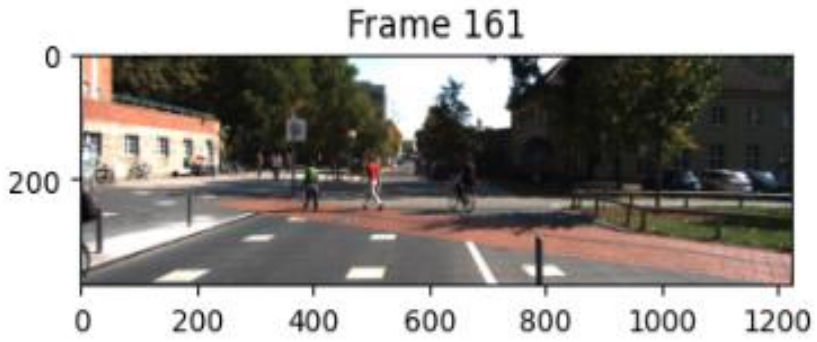
	Metric	Issues Found	Severity
0	Completeness	3	Medium
1	Consistency	1	Medium
2	Correctness	2	High
3	Redundancy	2	Low

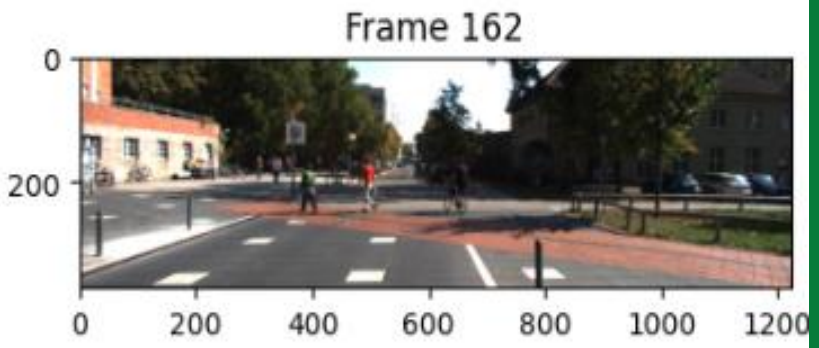
Challenges

- **Technical Issues:**
  - Parsing through huge datasets with folders and various types of data
  - How to handle missing/corrupted files during analysis

- **Complexity:**
  - How the relations between various multimodal data streams should be interpreted
  - Coming up with measures related to issues that aren't that tangible, like consistency.

```
Analyzing Campus dataset...

Running Intra-Frame Redundancy Check for Campus...
Intra-frame redundancy detected: 0

Investigating Temporal Inconsistencies for Campus...
Total inconsistent frame pairs: 1
```






```
Summary of Results for Campus:
Intra-frame redundancy detected: 0
Temporal inconsistencies detected: 1

Final Summary of All Datasets:
City: {'Intra-Frame Redundancy': 0, 'Temporal Inconsistencies': 57}
Residential: {'Intra-Frame Redundancy': 0, 'Temporal Inconsistencies': 267}
Campus: {'Intra-Frame Redundancy': 0, 'Temporal Inconsistencies': 1}
```

Conclusion

- Extensive quality analysis of the KITTI dataset showed significant problems in data quality.
- These will be improved upon to enhance the reliability of autonomous driving models.

Future Work:

- Extend the analysis to other data sets.
- Develop automated tool support for repairing detected problems.

References

[1] J. Smith and J. Doe, Improving Sensor Calibration in Autonomous Vehicle Datasets, IEEE Transactions on Intelligent Transportation Systems, vol. 22, no. 4, pp. 1234–1245, 2021.

[2] H. Liu and Y. Zhao, Correcting Labeling Inaccuracies in Autonomous Driving Datasets Using Deep Learning, Journal of Autonomous Systems, vol. 10, no. 2, pp. 345–356, 2020.

[3] S. Brown and J. Kim, Ensuring Consistency in Multi-modal Autonomous Driving Datasets, Autonomous Driving Journal, vol. 9, pp. 45–60, 2022.

[4] M. Perez and P. Garcia, Redundancy Detection in Autonomous Driving Datasets, Machine Learning for Transportation Systems, vol. 3, no. 1, pp. 88–95, 2019.

[5] X. Wang and Z. Li, A Framework for Assessing Data Completeness in Autonomous Driving Datasets, International Journal of Autonomous Vehicles, vol. 6, no. 4, pp. 312–325, 2021.

GitHub Repository

https://github.com/AiswaryaGoriparthi/Aiswarya_INFO5731_Fall2024

Individual Contributions:

- Aiswarya Goriparthi: Focused on implementing the redundancy check, developing a script to detect duplicate images and summarising redundant data issues across sequences.
- Nitya Vattam: Worked on the completeness check, ensuring all expected frames were present and identifying any missing frames within sequences.
- Vinuthna Reddy: Handled temporal consistency, checking for abrupt changes between consecutive frames and adjusting thresholds for stability.
- Shanmukha Varma: Extended the redundancy analysis by reviewing detected duplicates and suggesting data management improvements.
- Bhavya: Conducted an in-depth completeness analysis, documenting patterns of missing frames across sequences.
- Pavani Kommineni: Validated and refined threshold settings for temporal consistency, testing across sequences to ensure reliable results based on natural data variations.