Towards Object-Based Visual SLAM: A Revolution for Urban Tram

1 Introduction

Scope refinement. Following from the formulation of our research project (Part A): "Towards Object-Based Visual SLAM: A Revolution for Urban Tram", this document conducts the data analysis phase (Part B). Specifically, we now concentrate exclusively on **Stage 1: improving localisation accuracy in dynamic urban areas**. Therefore, our **response variable** (or target) is typically the presence, location of objects (e.g., cars, trucks, cyclists) within the images, which our object detection model aims to predict. The analytical findings in this report, which include initial results, figures, the analytical approach and visualisations, pave the way for further data preprocessing and modelling in Part C.

Refined research question. How can Big Data and deep learning-based object detection assist traditional SLAM systems in enhancing localisation accuracy for autonomous tram navigation in complex urban environments?

2 Exploratory Data Analysis (EDA)

In order to address our refined research target, we adopt and analyse two principal datasets: the BDD100K dataset (serving as the ***training set***), and the KITTI Object Detection Benchmark dataset (serving as the ***testing set***), accompanied by the corresponding training labels. The reason for our choice is as follows.

- **Data format:** For the object detection task in Stage 1 of our proposed plan, we leverage the YOLO model. As KITTI annotations are not compatible with the YOLO format [1], which requires us to seek a separate training dataset (BDD100K) to avoid extensive annotation conversion. This is due to the fact that BDD100K has bounding box annotations suitable for YOLO-based object detection [2].
- **Suitability:** The BDD100K dataset was specifically selected as a training set over other commonly used datasets, such as COCO, due to its strong emphasis on a wide variety of driving conditions. This dataset provides comprehensive coverage of urban driving situations (day, night, rain, fog), and comprehensive annotation of relevant moving objects (vehicles, pedestrians, traffic signs, and traffic lights), which we seek to further enhance the generalisation of our model to adapt to real-world driving tasks [2].

2.1 Initial Setup

Our primary dataset sources include:

- KITTI Object Detection Benchmark (***testing set***) [3].
- BDD100K (***training set***) [4].

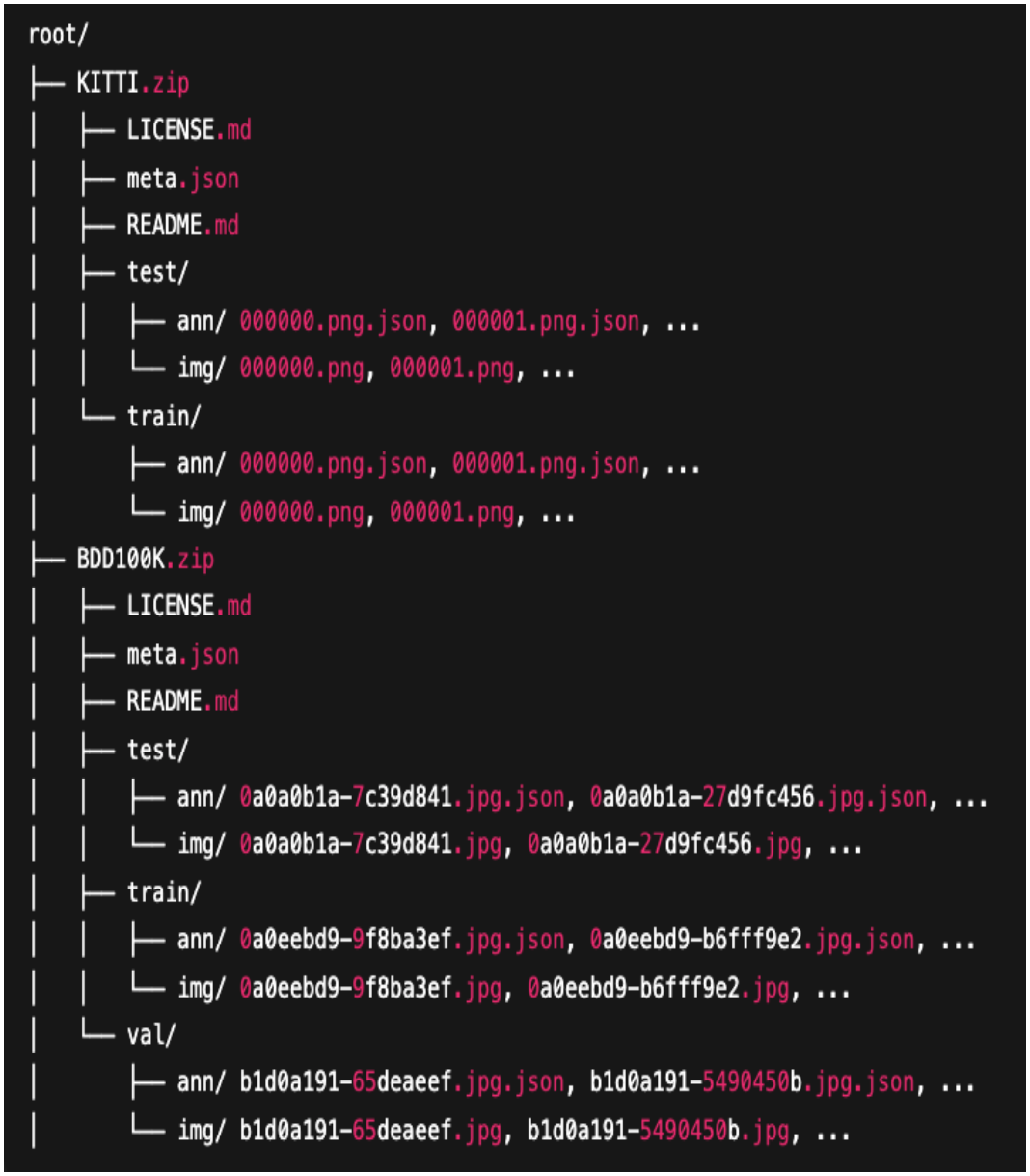


Figure 1: Directory structure diagram of the dataset

The Fig. 1 represents our dataset structure. Detailed configuration instructions for these requirements are provided in the `README.md` file of our GitHub repository.

2.2 Data Description

2.2.1 KITTI Object Detection Benchmark

The KITTI dataset, developed by the Karlsruhe Institute of Technology and Toyota Technological Institute, is a cornerstone of the computer vision community for the benchmarking of advanced algorithms in the field of autonomous driving [5]. In particular, KITTI object detection poses diverse

challenges emblematic of real-world driving scenarios and consists of **14,999** images with **51,865** labelled objects belonging to **9** different classes.

Key features.

- **type**: The object class.
- **bbox_xmin, bbox_ymin, bbox_xmax, bbox_ymax**: Pixel-based bounding box coordinates (the minimum and maximum x and y coordinates).
- **occluded**: This state indicates the object's degree of occlusion. The values include `fully visible, partly occluded, largely occluded`, which reflects varying levels of visibility. This is essential for SLAM systems to operate effectively.
- **observation_angle**: The observation angle expressed in radians.
- **dimensions_h, dimensions_w, dimensions_l**: the object's 3D measurements (height, width, and length in meters).
- **location_x, location_y, location_z**: 3D location of the object's centre in the camera coordinate system (meters).
- **rotation_y**: Rotation angle around the y-axis (radians).
- **bbox_width, bbox_height**: the derived pixel values for the width and height of the bounding box.

Visualisation.

1. **Object Class Counts (Univariate Analysis)**

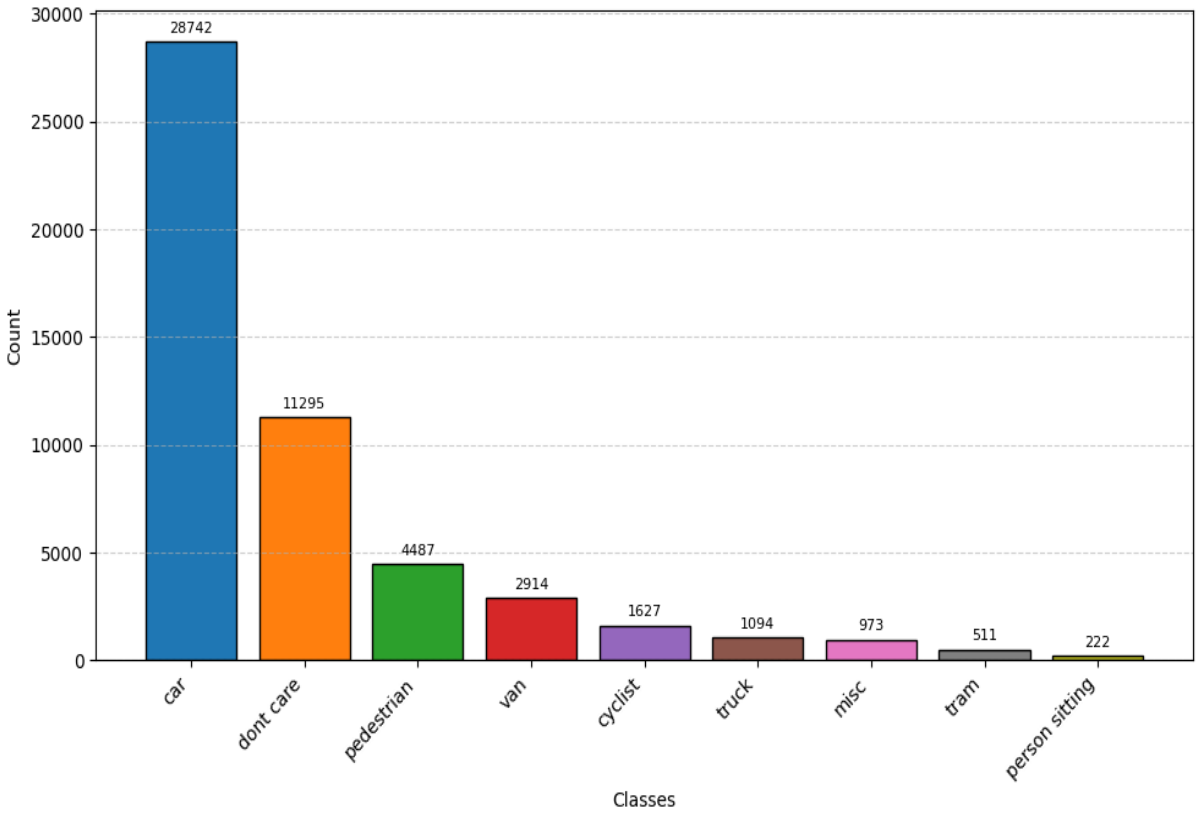


Figure 2: Distribution of Object Types in KITTI

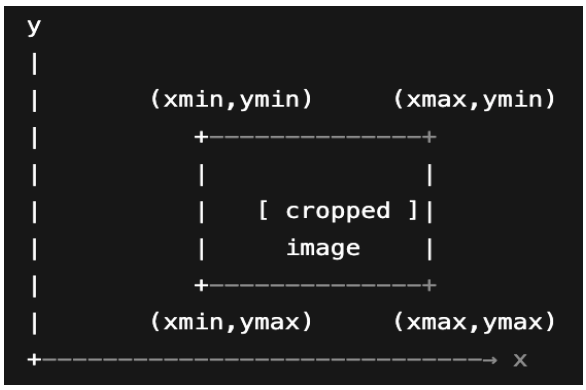
The bar chart in Fig. 2 illustrates the frequency of class types in the KITTI dataset. It can be clearly seen that the `Car` class dominates, while the others are minority classes. This is a typical scenario on the streets, which provides insight into why the prevalence of cars makes them ideal primary landmarks for the evaluation of the YOLO model in terms of improving localisation in dynamic urban environments [6].

2. **Sample Annotated Image and Bounding Box Insights (Bivariate Analysis)**

Figure 3: Sample Image with Bounding Box Annotations in KITTI

Images in the KITTI Object Detection dataset have bounding box annotations. As shown in Fig. 3, this is a sample image (`002126.png`) in the KITTI dataset. The image on the right side is the annotated version generated by extracting the bounding box coordinates and object labels from the annotation file corresponding to the original image on the left side. The vast majority of `Car` instances are labelled, while some boxes are annotated `dont_care`, signalling regions to be ignored during evaluation [1].

Insights about response variable. Objects are the entities (e.g., vehicles, pedestrians) annotated in images, serving as the dependent variables (response). In the sample image shown in Fig. 3, we utilised information from KITTI's `train` folder to visualise annotated images. Our decision to not use data from the `test` folder was due to its lack of detailed data and insufficient fields necessary for exploratory data analysis (EDA) and visualisation, aligning with the need for comprehensive attributes. As depicted in Fig. 3, pre-defined annotations of dispersed objects were employed for our analysis. The coordinate parameters for these object annotations were extracted from the `train` folder for visualisation purposes. Consequently, we utilised the `train` folder to support EDA, reserving the `test` folder for future model evaluation in Part C.



(a) Bounding Box Coordinates

(b) Cropped Image

Figure 4: Visualisation of representation of bounding box corner coordinates and cropped object (KITTI)

For a more in-depth examination of bounding boxes, we proceed by extracting an object within the annotated image from Fig. 3 and correlating it with its ground truth coordinates in Fig. 4a. From Fig. 4b, we can observe that the cropped region precisely encloses the visual extent of the object. Each of the corner corresponds to the symbolic coordinate in Fig. 4a. Specifically, the correlated coordinates are presented in the Tab. 1 below.

Corner	**Symbolic Coordinate**	**(x, y) Value**
Top-left	$(x_{\min}, y_{\min})$	(601, 183)
Top-right	$(x_{\max}, y_{\min})$	(675, 183)
Bottom-right	$(x_{\max}, y_{\max})$	(675, 254)
Bottom-left	$(x_{\min}, y_{\max})$	(601, 254)

Table 1: Bounding box corner coordinates with symbolic and numeric representation

Our comprehensive examination of summary statistics for bounding box coordinates is described in Fig. 5.

	bbox_xmin	bbox_ymin	bbox_xmax	bbox_ymax
count	51865.0	51865.0	51865.0	51865.0
mean	522.94	172.39	613.98	235.39
std	264.86	21.61	263.53	52.99
min	0.0	0.0	8.0	128.0
25%	366.0	166.0	454.0	199.0
50%	547.0	175.0	585.0	217.0
75%	684.0	182.0	744.0	253.0
max	1241.0	349.0	1241.0	375.0

Figure 5: Summary Statistics for Bounding Box Dimensions in KITTI

- According to this descriptive table, the bounding boxes are positioned vertically in the

lower half of the image, with mean values of:

$$\text{bbox}_{y_{\min}} \approx 172,$$
$$\text{bbox}_{y_{\max}} \approx 235,$$

and horizontally around the image's centre, with mean values of:

$$\text{bbox}_{x_{\min}} \approx 523,$$
$$\text{bbox}_{x_{\max}} \approx 614.$$

- The high `std` in x-coordinates (≈ 264) indicates objects are dispersed across the image width, representing a variety of urban settings.
- Objects tend to extend further down the image, as indicated by the moderate `std` in y-coordinates (≈ 53).

3. Spatial Heatmaps (Bivariate Analysis)

As shown in the plot of Fig. 6, this spatial heatmap depicts the density of top nine KITTI objects across images. These visualisations provide insights into the most probable and rare object locations on the image. It helps analyse objects' placements in a dataset. The `Car` instances cluster mostly at the image centre, with the bright horizontal band spreads around x-coordinates $\approx$ 500-700px, and y-coordinates $\approx$ 170-220px. This suggests that cars mostly appear in the middle-lower regions of the image, reflecting their real-world camera viewpoint. While other classes, such as `cyclists`, `pedestrians`, and `person sitting` exhibit darker, less intense clusters, mirroring their presence near the sidewalks. Overall, this distribution accentuates the structured layout in urban corridors, requiring the need for prioritising lower and central regions through region-of-interest (ROI) cropping or adaptive anchor-box initialisation. This has been proved to boost YOLO efficiency [6, 7].

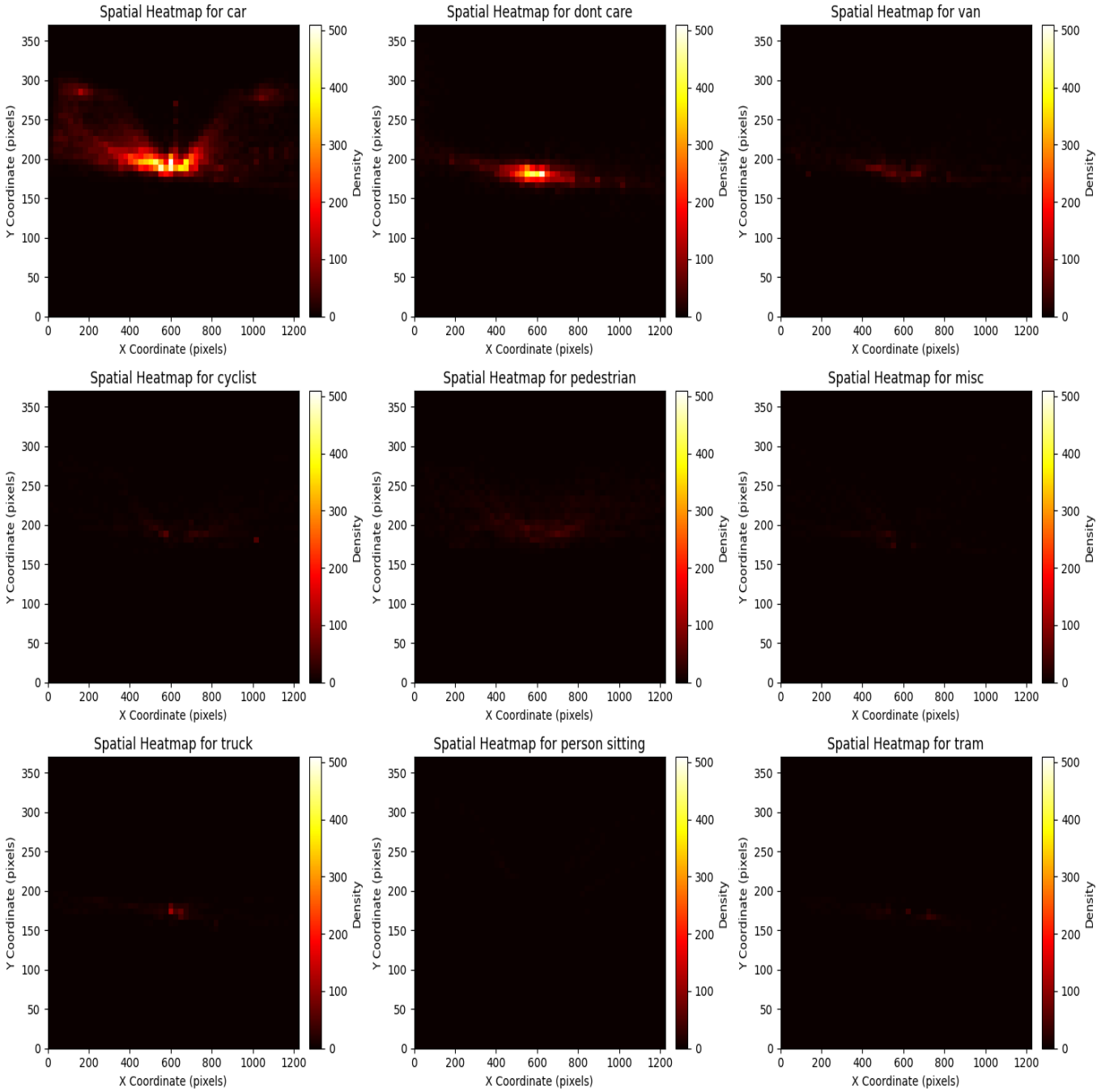


Figure 6: Spatial Heatmaps of Object Classes in KITTI

4. Co-Occurrence Matrix (Multivariate Analysis)

Co-occurrence matrix is an extremely valuable tool that shows the images for every pair of classes: how often two categories appear at the same time (e.g., the same image or frame). In Fig. 7, we visualise the co-occurrence matrix for the KITTI dataset displays the frequency of object type pairs across 8 categories (note: the `lane` category was excluded to focus the analysis on dynamic objects).

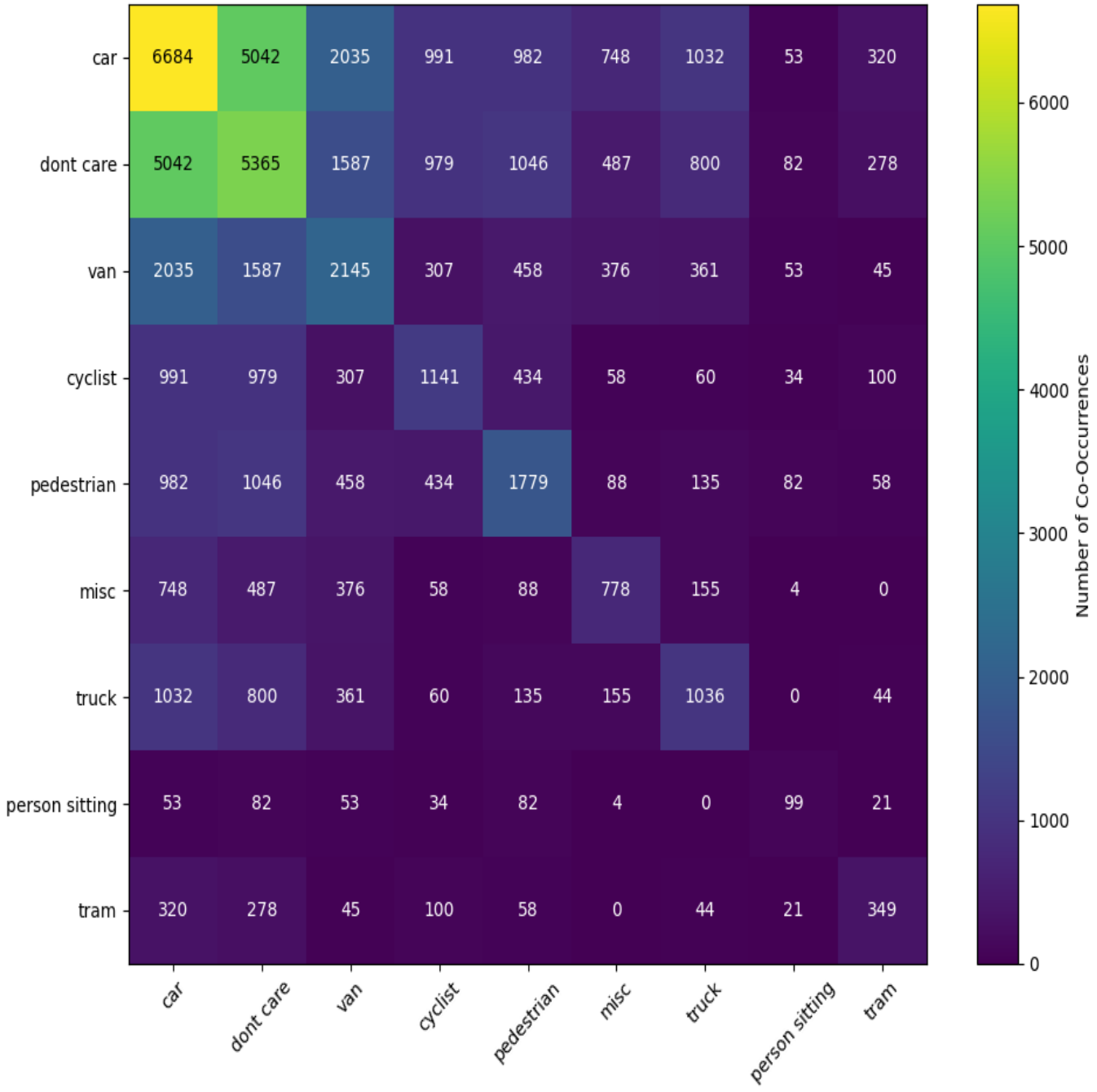


Figure 7: Co-Occurrence Matrix of Object Types in KITTI

Most important statistics:

- **Highest Co-Occurrence:** The pair `car - car` dominate (**6,684** counts), suggesting dense vehicle clusters, key landmarks for SLAM mapping [1].
- **Notable Cross-Category:** The `car - pedestrian` co-occurrences (**982** counts) indicate mixture of vehicle and pedestrian, enabling the system to predict and track distinct objects, hence support the feature association in crowded areas [8].
- **Low Co-Occurrence:** We observe the `tram - tram` pairing has low values (**349** counts), which means trams are less clustered. Some pairs have zero values, i.e., `tram - truck`.

The aforementioned statistics support predictive modelling, where frequent object pairs guide real-time mapping and reduce localisation errors in dynamic settings.

2.2.2 BDD100K

The BDD100K dataset, which stands for Berkeley Deep Drive Dataset, is a dataset for instance segmentation, semantic segmentation, object detection, and identification tasks. It is used in the automotive industry. The dataset consists of **100,000** images with **2,221,128** labelled objects belonging to **12** different classes. Images 100K dataset have pixel-level instance segmentation annotations. Due to the nature of the instance segmentation task, it can be automatically transformed into a semantic segmentation (only one mask for every class) or object detection (bounding boxes for every object) tasks [4].

Key features.

- **type**: The object class.
- **bbox_xmin, bbox_ymin, bbox_xmax, bbox_ymax**: Pixel-based bouding box coordinates (the minimum and maximum x and y coordinates).
- **occluded**: This state indicates the object's degree of occlusion. The values include boolean attribute: `True, False`, indicating whether the object is occluded (partially hidden by another object).
- **truncated**: This state indicates the extension of object beyond the image boundary and is therefore partially cut off. The values include boolean attribute: `True, False`, indicating whether the object is truncated (partially outside the image frame).

Visualisation.

1. **Object Class Counts (Univariate Analysis)**

   Fig. 8 depicts the bar chart, representing frequency of object types in the BDD100K dataset. We can observe that this dataset exhibits a more balanced distribution compared to KITTI (as shown in Fig. 2). While cars remain prevalent (**713,211**), substantial numbers of `lanes, traffic signs, traffic lights` and other features also present. This insight strengthens our choice of BDD100K for a balanced training set for YOLO model capable of detecting a wide range of objects; and is therefore critical for urban tram navigation where interactions with diverse elements are frequent [9].

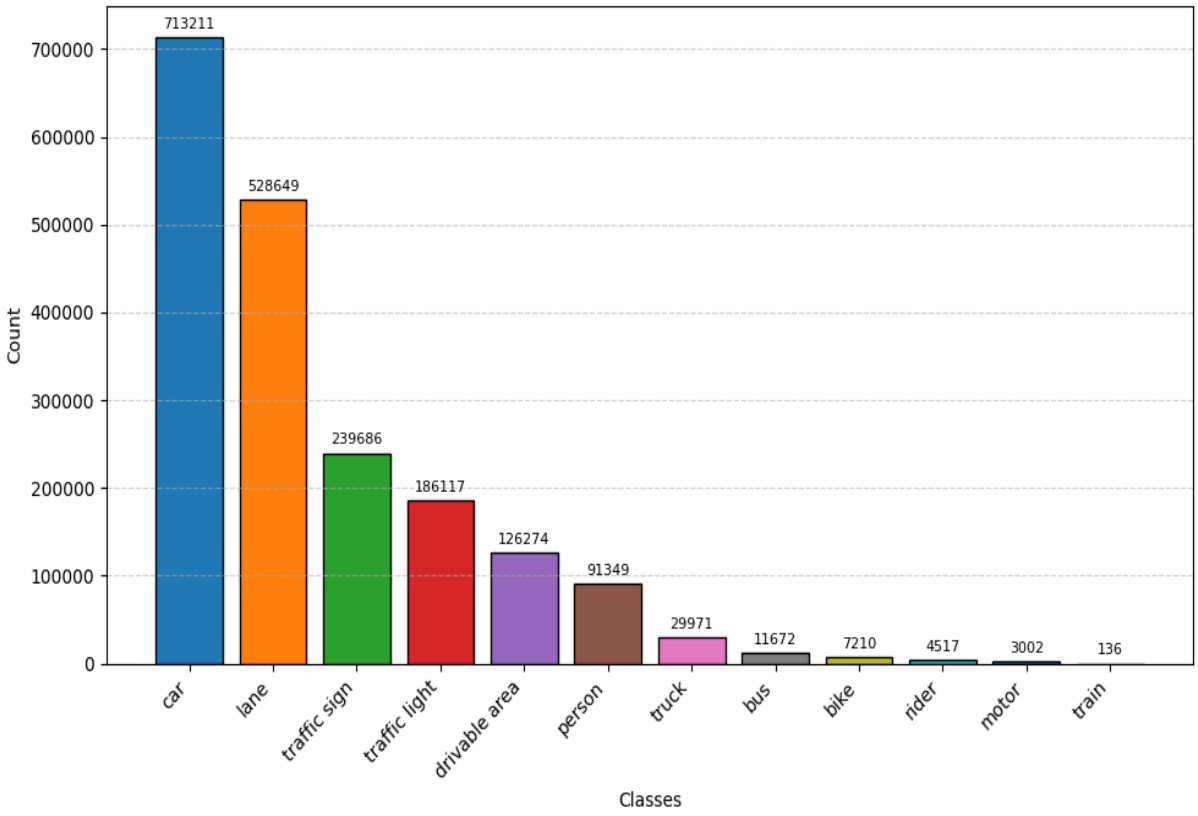


Figure 8: Distribution of Object Types in BDD100K

2. Sample Annotated Image and Bounding Box Insights (Bivariate Analysis)

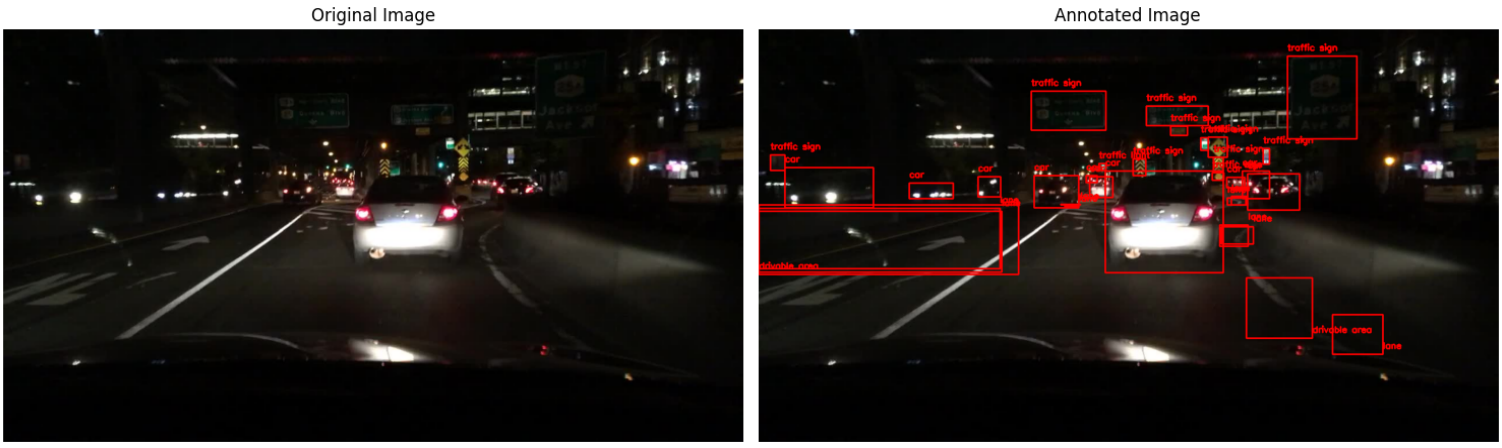
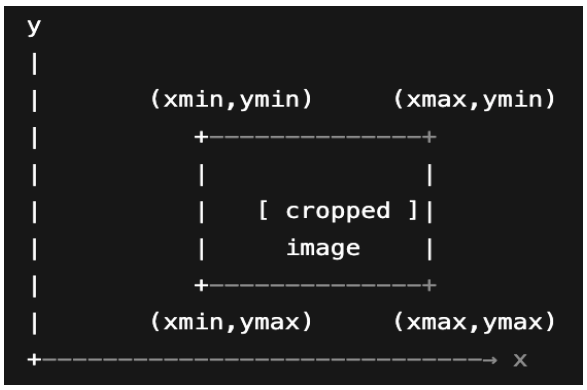
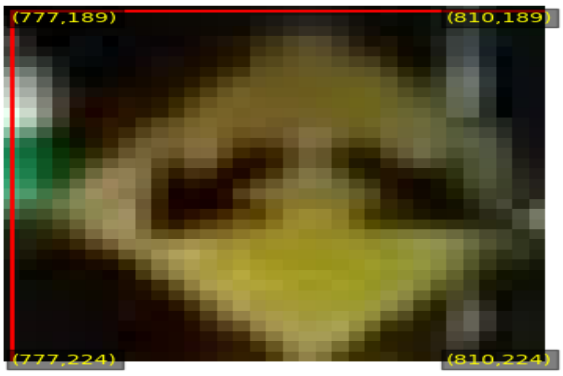


Figure 9: Sample Image with Bounding Box Annotations in BDD100K

Images in the BDD100K dataset also contain bounding box annotations, similar to KITTI. Fig. 9 showcases a sample image (`0a006b7b-c22407a2.jpg`) from `train` folder of BDD100K dataset. However, as opposed to the KITTI sample image from Fig. 3. This particular sample was captured during night-time city with multiple annotated objects (i.e., cars, lane, traffic sign, drivable area) under challenging lighting conditions, while KITTI sample image typically feature fewer object annotations, primarily focusing on cars. Although the inherent variability poses challenges but it provides generalisability to a more robust object detection model, which is crucial for accurate localisation in complex urban environments [2].



(a) Bounding Box Coordinates



(b) Cropped Image

Figure 10: Visualisation of representation of bounding box corner coordinates and cropped object (BDD100K)

Similar to KITTI bounding box analysis, we present the corresponding coordinates in Fig. 10. Each of the corner matches with the symbolic coordinate in Fig. 10a. Specifically, the correlated coordinates are presented in the Tab. 2 below.

Corner	**Symbolic Coordinate**	**(x, y) Value**
Top-left	$(x_{\min}, y_{\min})$	(777, 189)
Top-right	$(x_{\max}, y_{\min})$	(810, 189)
Bottom-right	$(x_{\max}, y_{\max})$	(810, 224)
Bottom-left	$(x_{\min}, y_{\max})$	(777, 224)

Table 2: Bounding box corner coordinates with symbolic and numeric representation

Our comprehensive examination of summary statistics for bounding box coordinates is described in Fig. 11.

	bbox_xmin	bbox_ymin	bbox_xmax	bbox_ymax
count	1941794.0	1941794.0	1941794.0	1941794.0
mean	559.59	345.72	611.45	390.71
std	303.55	116.2	326.26	122.99
min	0.0	0.0	0.0	0.0
25%	358.0	276.0	387.0	312.0
50%	560.0	340.0	594.0	381.0
75%	748.0	409.0	815.0	461.0
max	1279.0	719.0	1279.0	719.0

Figure 11: Summary Statistics for Bounding Box Dimensions in BDD100K

- The significantly higher count values (**1,941,794**) for all bounding box coordinate attributes in BDD100K, outnumbering KITTI (**51,865**), which makes it ideal for BDD100K to serve as training set. In addition, this finding signifies BDD100K's extensive coverage

and diversity, improving its reliability for robust training of object detection models in diverse urban driving conditions [2].

- According to the descriptive figures, the bounding boxes in the BDD100K dataset are located vertically around the lower-middle region of the image, with mean values of:

$$\text{bbox}_{y_{\min}} \approx 346,$$
$$\text{bbox}_{y_{\max}} \approx 391,$$

and horizontally around the image's centre, with mean values of:

$$\text{bbox}_{x_{\min}} \approx 560,$$
$$\text{bbox}_{x_{\max}} \approx 611.$$

This suggests that the BDD100K shares the same central-horizontal cluster pattern with KITTI, but is slightly higher vertically.

- All four xy-coordinate features show a minimum value of 0, suggesting the presence of objects at or beyond the boundaries of the image, this also adheres to the truncated characteristic of the dataset, as realistic situations arise when objects were only partially visible within the camera frames, as in the case of images from driving scenarios [2].

3. Spatial Heatmaps (Bivariate Analysis)

Similar to KITTI, the spatial heatmaps in Fig. 12 reveal unique spatial patterns of top nine object classes in BDD100K across images. The `car` instances show high density in the lower centre (x-coordinates $\approx$ 500–800 px, and y-coordinates $\approx$ 300–400 px). Other categories such as `traffic sign` is more peripheral, often appearing on the sides of image frames.

Again, as previously mentioned in the KITTI section, this distribution accentuates the structured layout in urban corridors, requiring prioritisation of lower and central regions through region-of-interest (ROI) cropping or adaptive anchor-box initialisation. In the context of BDD100K, which serves as our training set, the spatial heatmap investigation lead us to two proposed preprocessing options:

- **ROI cropping**: Trimming each frame to the lower–central band where cars and lanes most frequently occur can cut pixel throughput while preserving detection accuracy, a technique empirically shown to accelerate inference without accuracy loss [10].
- **Anchor-box configuration**: Single-stage detectors such as YOLOv8 perform automatic anchor optimisation during training by learning anchor shapes directly from the dataset bounding-box distribution, reducing the need for manual k-means clustering [11].

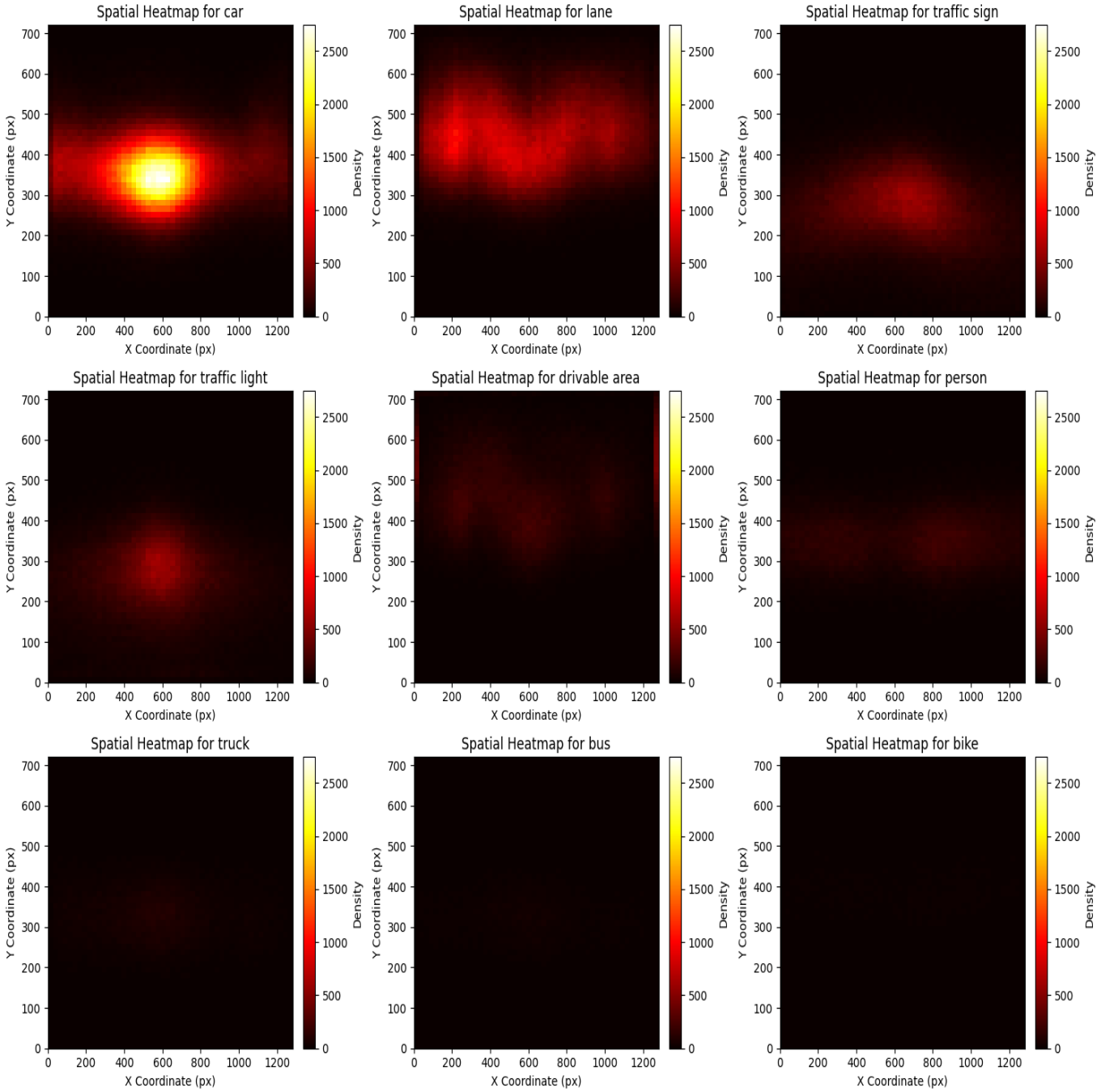


Figure 12: Spatial Heatmaps of Object Classes in BDD100K

4. Co-Occurrence Matrix (Multivariate Analysis)

Likewise, the co-occurrence matrix in Fig. 13 illustrates how frequently pairs of object categories appear together in the BDD100K dataset.

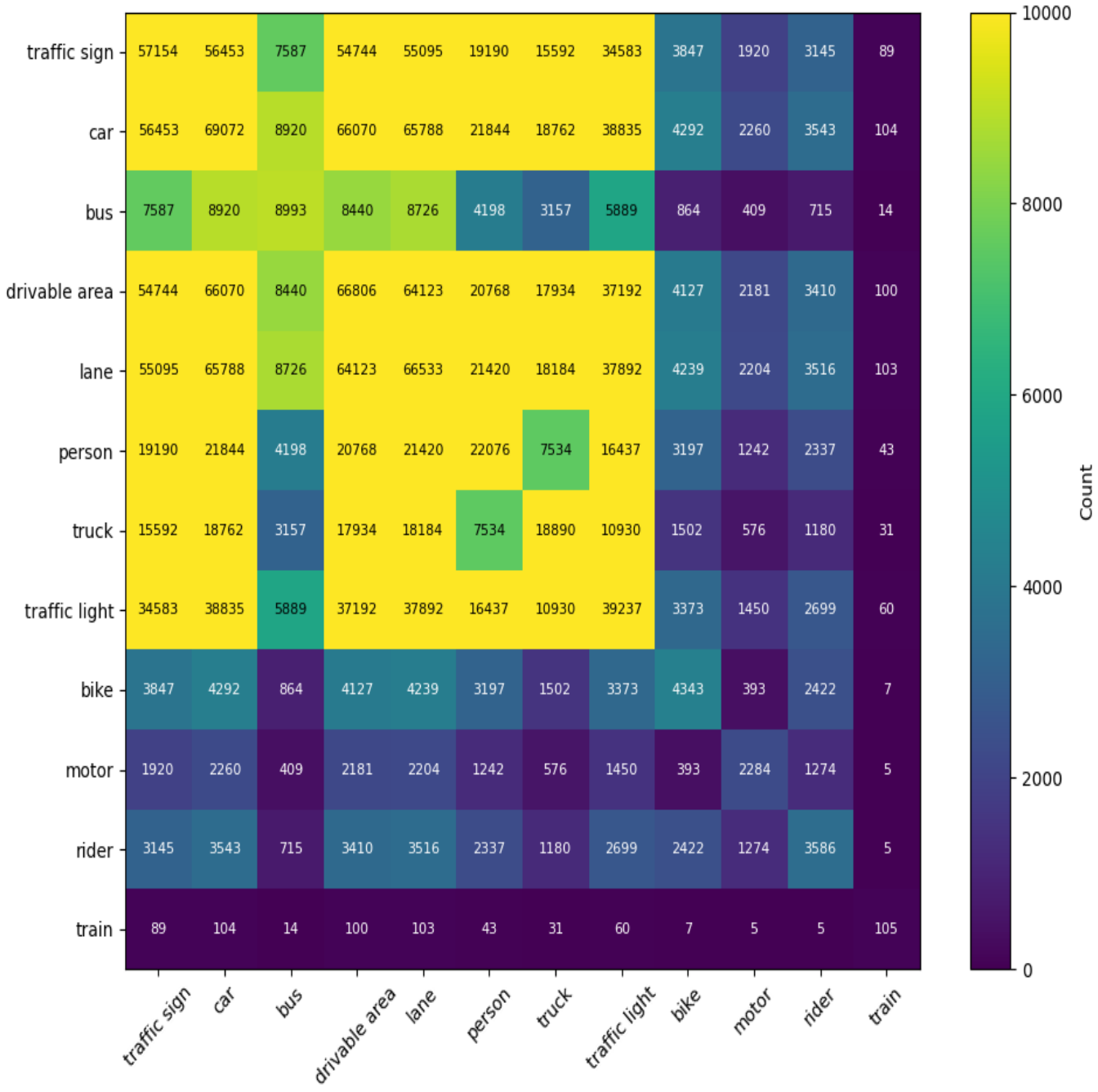


Figure 13: Co-Occurrence Matrix of Object Types in BDD100K

Most important statistics:

- **Highest Co-Occurrence:** The pair `car - car` still prevalent (**6,684** counts), reflecting significantly high vehicle clusters. This `car` pairing makes them crucial landmarks for effective feature tracking and object-based localisation.
- **Notable Cross-Category:** The `car - traffic sign` co-occurrences (**56,453** counts) and `car - lane` indicate a high variety of features, encompassing both dynamic (e.g., vehicles) and static objects (e.g., traffic sign, lane). This mixture offers good opportunity to train a robust model in visual SLAM applications.
- **Low Co-Occurrence:** We observe the several pairings have substantially lower values

(e.g., `train - bus` with only **14** counts, and `train - traffic sign` with **89** counts), which suggests sparse interactions of several vehicles within urban settings.

2.2.3 Aspect Ratio Comparison of KITTI and BDD100K for Object Detection

Aspect ratio is the proportional relationship between an image's width and height, expressed as width:height [12]. Upon examining the configuration metadata and annotation files provided with each dataset, we found and calculated that KITTI and BDD100K datasets have image aspect ratios of approximately `3.3:1` and `16:9` (based on standard annotations), respectively. This affects YOLO's anchor box design and field of view, necessitating dataset-specific preprocessing to optimise localisation accuracy in SLAM applications [13, 14].

3 Conclusion

In this work, we have thoroughly gone through two principal datasets, KITTI and BDD100K, to address our refined research question: enhancing localisation accuracy for autonomous tram navigation using Big Data and deep learning-based object detection. Given that KITTI and BDD100K are well-established datasets with curated annotations; however, using different datasets for training and testing presents some important considerations regarding data cleaning and outlier removal. Therefore, this will require domain adaptation technique for further preprocessing steps. However, in the scope of this phase, we concentrated exclusively on analysing the characteristics of both datasets only, as our project—unlike other traditional machine learning pipelines—requires an in-depth examination of two distinct datasets. Detailed steps relating to outlier handling, data cleaning, and further data preprocessing will be consolidated and elaborated in Part C, where we will define and implement the final preparation procedures before modelling and evaluation.

4 Code Repository

The code for this document is available as an open-source project on GitHub at: GitHub Repository: Object-based Visual SLAM. The repository also includes a `README.md` file that provides detailed instructions for setting up the environment, installing dependencies, and reproducing the analyses and visualisations described in this report.

References

[1] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3354–3361, 2012. Accessed: 16 June 2025.

[2] Fisher Yu, Haofeng Chen, Xin Wang, Weichao Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2636–2645, 2020. Accessed: 16 June 2025.

[3] Dataset Ninja. Visualization tools for kitti object detection dataset, 2025. Accessed: 20 June 2025.

[4] Dataset Ninja. Visualization tools for bdd100k: Images 100k dataset, 2025. Accessed: 20 June 2025.

[5] Abhishek Gupta et al. Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues. *Array*, 10:100057, 2021. Accessed: 30 June 2025.

[6] Y. Gefan and L. Yuchi. Object Detection in the KITTI Dataset using YOLO and Faster R-CNN. *International Journal of Advanced Computer Science and Applications*, 10(5):257–263, 2019. Accessed: 30 June 2025.

[7] Holger Caesar et al. nuScenes: a multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11621–11631, 2020. Accessed: 02 July 2025.

[8] Raul Mur-Artal, Juan M. M. Montiel, and Juan D. Tardós. ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015. Accessed: 02 July 2025.

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016. Accessed: 03 July 2025.

[10] Jinwoo Jeon and Jaehoon Lee. A study on the yolov7 object detection algorithm using transfer learning. *Applied Sciences*, 15(10):5328, 2025. Accessed: 03 July 2025.

[11] Ultralytics. YOLOv8 Documentation, 2024. Accessed: 03 July 2025.

[12] Richard Szeliski. *Computer Vision: Algorithms and Applications*. Springer, Cham, 2 edition, 2022. p. 23. Accessed: 04 July 2025.

[13] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. pp. 4–5. Accessed: 04 July 2025.

[14] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, Jose Neira, Ian Reid, and John J. Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics*, 32(6):1309–1332, 2016. Accessed: 04 July 2025.