



MARMARA UNIVERSITY

FACULTY OF ENGINEERING

CSE4288

Introduction to Machine Learning

TERM PROJECT

Data Preprocessing and EDA

Group: 2

Table of Contents

1. Dataset	3
2. Data Cleaning and Preprocessing	3
3. Exploratory Data Analysis (EDA)	4
4. Feature Identification and Engineering	5

1. Dataset

For this project, we are using BDD100K dataset, which is a public dataset including images and videos taken from car cameras. This multimedia consists of daily traffic through the city: vehicles driving, pedestrians walking etc.

We are responsible for detecting the crosswalks in images. Therefore, only images are retrieved as the raw dataset.

2. Data Cleaning and Preprocessing

The dataset needs to be cleaned to begin preprocessing. In Section 1, it is mentioned that images consist of various daily vehicle and pedestrian traffic. However, we are only responsible with the images that includes crosswalk labelling for detecting the crosswalks. Moreover, among these images, there are images that are not labelled at all. Firstly, we cleaned images without labels. Then, we cleaned the images that do not include crosswalks.

After successfully cleaning the raw dataset, the second step is data preprocessing, where we need to process data in a way that can be the input for YOLOv8 training.

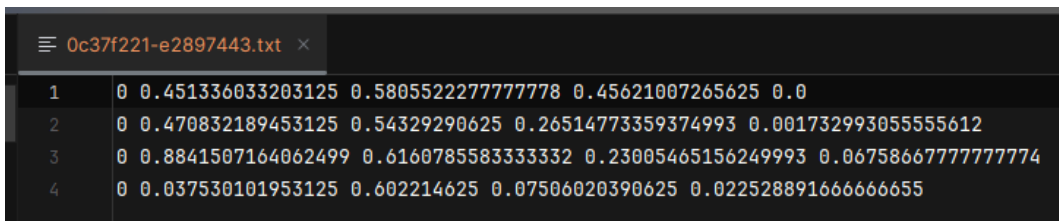
The crosswalk label includes the vertices of the 2d polygon, indicating the location of the object inside the image. In Figure 1, an example of crosswalk labelling is shown.

```
"category": "crosswalk",  
"poly2d": [  
  {  
    "vertices": [  
      [  
        490.49528,  
        419.568264  
      ],  
      [  
        576.406877,  
        416.571348  
      ]  
    ],  
    "types": "LL",  
    "closed": false  
  }  
]
```

Figure 1: Crosswalk labelling of an image

The labels are given as 2D polygons using vertices. However, YOLO's bounding boxes method is compatible with rectangular formed coordination. First thing to do is to convert these polygon vertices into rectangular form. This rectangle is defined by rightmost, leftmost, topmost and the bottommost points of the polygon.

After the conversion, the information about the image needed to be written in YOLO format, including the class ID, width and height of the crosswalk. In this project, the class ID of crosswalk is set as 0. Additionally, after calculating the width and height from the rectangular coordinates, these values should be normalized and recorded in txt file for each image. An example for YOLO format txt file is given in Figure 2.



```
0c37f221-e2897443.txt x
1 0 0.451336033203125 0.5805522277777778 0.45621007265625 0.0
2 0 0.470832189453125 0.54329290625 0.26514773359374993 0.001732993055555612
3 0 0.8841507164062499 0.616078558333332 0.23005465156249993 0.06758667777777774
4 0 0.037530101953125 0.602214625 0.07506020390625 0.022528891666666655
```

Figure 2: YOLO format txt file of training image

3. Exploratory Data Analysis (EDA)

Table 1: Analysis of the raw dataset

Total number of images in dataset	7000
Total number of images with crosswalk	4896
Total number of crosswalk annotations	1103
Average crosswalk number per image	4
Maximum crosswalk number per image	17
Minimum crosswalk number per image	1

The dataset for this project includes 7,000 images, of which 4,896 contain crosswalk labels. A total of 1,103 crosswalks are annotated, with an average of 4 crosswalks per image. Some images have up to 17 crosswalks, while others have just 1, highlighting the need for careful preprocessing and robust model design. Images with multiple crosswalks provide valuable training examples, but the variability in crosswalk density poses challenges for accurate detection in complex scenes. To streamline training, 2,104 images without crosswalks were excluded, ensuring the model focuses on relevant features. This diversity in crosswalk distribution is key to training a YOLOv8 model capable of generalizing well to new images.

4. Feature Identification and Engineering

As mentioned in Section 1, raw dataset consists of various labels. In Figure 3, labels are described.

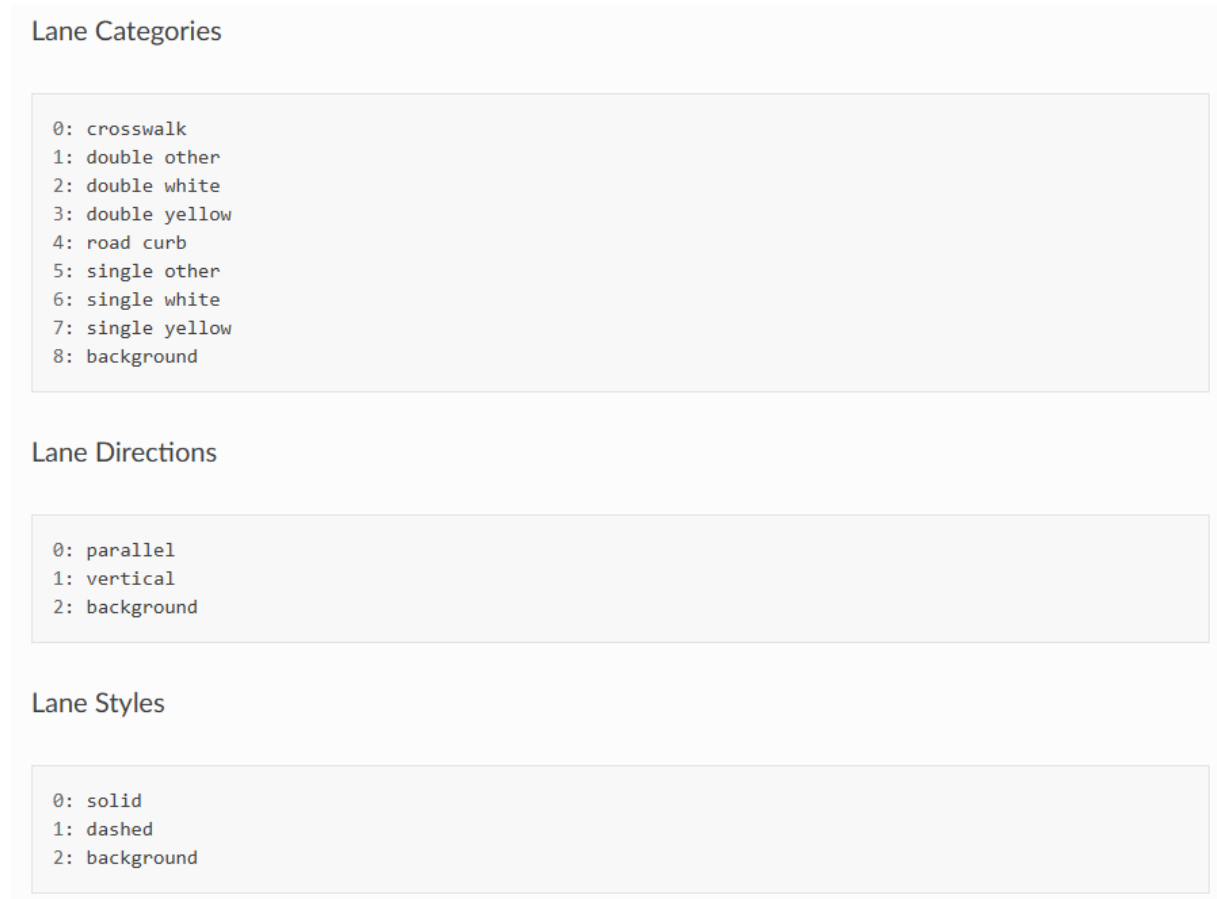


Figure 3: Lane Marking

The raw dataset includes labels for multiple categories, such as lanes, road curbs, and other markings. Since the objective of this project is to detect crosswalks, all other annotations were excluded. This filtering process ensured that only the "crosswalk" category (class ID 0) was included in the final dataset. This reduces noise and focuses the model on learning features specific to crosswalk detection.

As the dataset contains only one class, there is no direct issue with class imbalance. However, ensuring diverse examples of crosswalks such as different lighting conditions, angles, and environments was prioritized inside the dataset. This diversity helps the model generalize better during inference, particularly for unseen scenarios.

BDD100K provides different types of datasets for training, validating and testing. Therefore, the training dataset is not separated currently. All the images are used for training purposes inside the training dataset. After for validation and testing phases, associated dataset will be used.

About the cleaning of the training dataset, images without crosswalk annotations were removed, as they do not contribute to the training process. Additionally, all bounding box coordinates were normalized to the image dimensions during preprocessing. This normalization step ensures that the model can handle images of varying sizes consistently without requiring resizing during training. Additionally, the pixel values of the images were normalized to improve the model's numerical stability during optimization.

These steps ensure that the dataset is clean, relevant, and compatible with YOLOv8's requirements and ready for model development.

5. References

[1] Fisher Yu. 2024. "BDD100K Documentation". Date Accessed: 25.11.2024.

(<https://doc.bdd100k.com/index.html>)

[2] Ultralytics, 2024. "Ultralytics YOLO Documentation". Date Accessed: 25.11.2024.

(<https://docs.ultralytics.com/tr>)