Visual Computing Magazine

# Road Object Detection and distance estimation using depth monocular model

Baroud Yasmine, Bourzam Saad Allah, LAICHE Nacera, USTHB

**Abstract**.
Accurate detection and recognition of road objects, especially small objects, are crucial for autonomous driving. In this article, we have designed two approaches to solve the problem of detecting small road objects and estimating the distance between the detected objects and the vehicle's camera. The first approach is based on object detection without segmentation, where we optimized the YOLOv5s model. The second approach involves object detection with segmentation using the Mask R-CNN model with Detectron2.
Furthermore, we developed two different models for estimating the absolute distance of the detected objects. The first model utilizes the object detection of optimized model along with information from a monocular depth map using MiDaS. The second model is a deep learning instance segmentation model that extracts specific information from the masks of the detected objects and utilizes relative distances generated by MiDaS to estimate the absolute distance (m).


Figure 1. Example of result of road detection and recognition for a BDD10K-MV test image obtained with our optimized model.

## 1. Introduction

Real-time road object detection is critical for scene recognition and safe navigation of autonomous devices in natural environments. Furthermore, precise distance estimation is essential for advanced autonomous driving systems to provide safety features such as adaptive cruise control and collision avoidance. While radars and lidars can provide distance information, they are either costly or less accurate than image sensors when it comes to object information. However, detecting small objects and objects located at large distances in road images remains challenging [1], particularly using light models.

Additionally, distance estimation between the detected objects and vehicle's camera is very challenging using 2D monocular camera. In this article, we have proposed two methods to solve the problem of detecting small road objects and estimating the distance between the detected objects and the vehicle's camera using a single image.



# Road Object Detection and distance estimation using depth monocular model

Baroud Yasmine, Bourzam Saad Allah, LAICHE Nacera, USTHB

## 2. Method

We proposed two methods for designing a system that detects and recognizes road objects and estimates the distances between the detected objects and the camera using a single image.
Our design consists of three main parts:

### *2.1. Object Detection and Recognition:*

- Method 1: We utilized our proposed model, based on a combination of YOLOv5s [2] and Swin transformer [3], for object detection without segmentation.
- Method 2: We employed the Mask R-CNN model [4] with Detectron2 [5] for segmentation-based object detection. The goal is to compare performance and determine the optimal execution speed for our real-time distance estimation system.

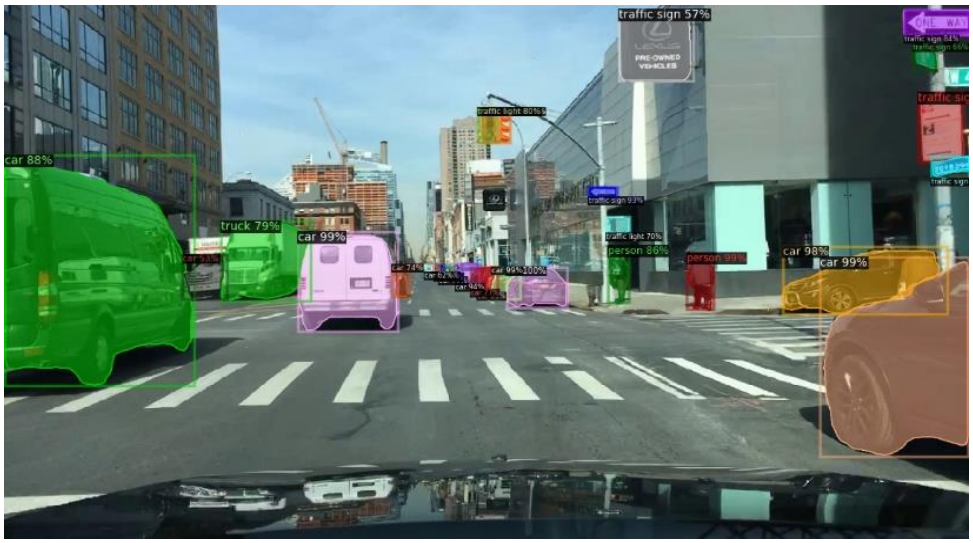


Figure 2. Example of result of road detection, segmentation and recognition for a BDD10K-MV test image obtained with our model trained Mask R-CNN with Detectron2..

### *2.2. Distance Estimation:*

- We developed two different models for estimating the absolute distance of detected objects. The first model combines the object detection results with information from a monocular depth map generated by the MiDaS model. The second model utilizes deep learning instance segmentation (Mask R-CNN with Detectron2), specific mask information, and information from MiDaS [6].

# Road Object Detection and distance estimation using depth monocular model

Baroud Yasmine, Bourzam Saad Allah, LAICHE Nacera, USTHB

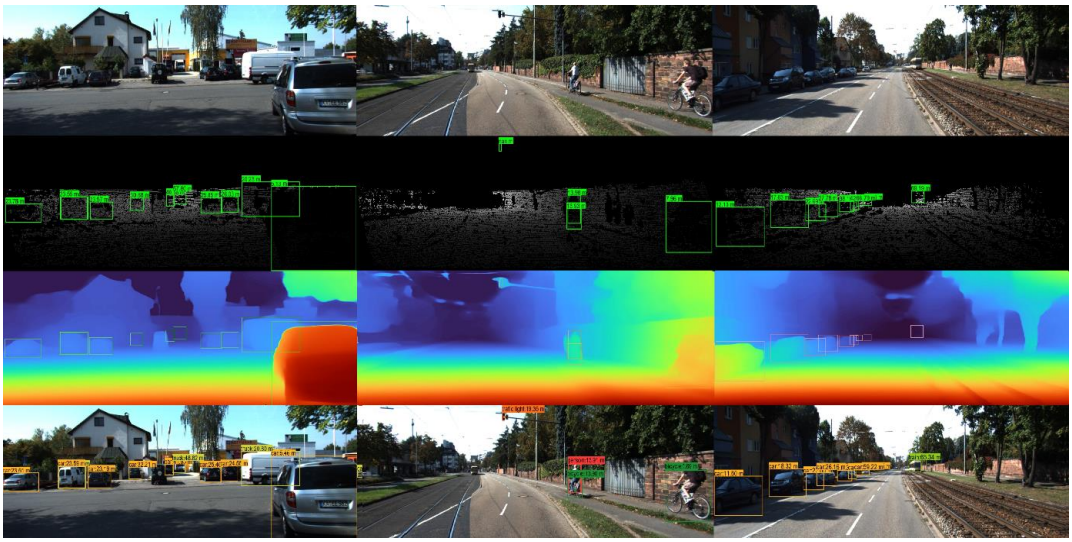

Figure 3. Examples of estimating distance using Mask R-CNN with Deetectron2 and MiDaS model on KITTI dataset.

The process involves running the two models in parallel. The MiDaS model predicts the relative depth map of the input image, while the proposed model locates and classifies road objects within the image. The location results of each object defined by bounding boxes obtained are overlaid on the estimated relative depth image. The relevant (relative) distance of an object is calculated by averaging the relative distance of all pixels not null within the defined bounding box in the MiDaS estimated depth map. To convert the relative distance (REV) of an object into real distance (ABS), the real distance of the objects in the images is estimated using LIDAR point clouds provided as ground truths in the KITTI Raw dataset. Then, the relationship between the real (absolute) distance and the relative distance is calculated using a quadratic mathematical formula that minimizes the error between the actual distance and the relative distance.

For the second approach mentioned above, we followed the same steps, but instead of considering the points inside the bounding box generated by the proposed model, we considered the points inside the object mask detected and segmented by the Mask R-CNN model with Detectron2.

### 2.3. Data augmentation:

To address the limited availability of data for certain road object classes, we proposed two methods for data augmentation. The first method involves applying data augmentation techniques such as photometric distortion, geometric distortion, etc, making the models more robust

# Road Object Detection and distance estimation using depth monocular model

Baroud Yasmine, Bourzam Saad Allah, LAICHE Nacera, USTHB


Figure 4 Example of Road object detection and recognition result for an image of Algiers obtained by our model optimized.

against images from different environments and lighting conditions. The second method aims to increase the number of instances for classes with low representation in the BDD10K dataset, strengthening the object representation and achieving better balance among classes. We refer to this new dataset as BDD10K-MV (BDD10K-Mapillary Vistas) in our article.

### 3. Experimental results

In this section, we present the results of our experiments conducted to evaluate the performance of the proposed approaches. Firstly, we evaluated the proposed model through various experiments using the KITTI dataset toward six different road objects (person, bicycle, car, motorcycle, bus, train, truck) and our newly created dataset BDD10K-MV with nine road object (Person, Traffic Sign, Traffic light, Car, Truck, Motorcycle, Train, Bus Bicycle). The evaluation metrics focused on mean average precision (Map50%) and revealed that the proposed model achieved an impressive performance with a Map50% score of 95% on the KITTI dataset and 56.4% on the BDD10K-MV dataset. Compared to YOLOv5s, our approach improves mAP by 1.5% and 3.5% on the KITTI and BDD10K-MV datasets respectively, enabling higher accuracy in detecting small objects road images and objects located at a large distance.

Comparatively, we assessed the Mask R-CNN model with Detectron2, which achieved an average precision (AP50%) of 56.88% on the BDD10K-MV dataset.

The estimation of absolute distance was evaluated using all images from the KITTI Raw dataset, and the proposed models demonstrated their effectiveness, which makes the solution highly competitive with existing approaches. The choice of model ultimately depends on specific needs and objectives. For our focus on precise distance estimation and real-time detection in autonomous vehicles, we selected the proposed model of detection and recognition road objects and MiDaS variant of dpt_swin2_l384 as the preferred detection and recognition with distance estimation model.

# Road Object Detection and distance estimation using depth monocular model

Baroud Yasmine, Bourzam Saad Allah, LAICHE Nacera, USTHB


Figure 5. Detection, recognition and segmentation results for road objects in a BDD10K test image obtained by the Mask R-CNN model with Detectron2


Figure 6. Road object detection and recognition results on two images from the UAVDT UAVDT_benchmark drone dataset obtained with our optimized model, showing the effectiveness of our model in detecting small objects.

**References**

[1] Armin Masoumian et al. “Absolute Distance Prediction Based on Deep Learning Object Detection and Monocular Depth Estimation Models”. In: (oct. 2021). url : https://arxiv.org/abs/2111.01715
[2] Ultralytics. “You Only Look Once : Unified, Real-Time Object Detection”. In : (2022). url : https://github.com/ultralytics/yolov5/issues/6998.
[3] Ze Liu et al. “Swin Transformer : Hierarchical Vision Transformer using Shifted Windows”. In : (2021). arXiv : 2103.14030 [cs.CV].
[4] Kaiming He et al. “Mask R-CNN”. In : (2018). arXiv : 1703.06870 [cs.CV].
[5] Meta AI. Detectron2 : A PyTorch-based modular object detection library. consulté en avril 2023. url : https://youtu.be/egs0XN-xjA0
[6] René Ranftl et al. “Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer”. In : (2020). arXiv : 1907.01341 [cs.CV].