ROAD OBJECT DETECTION
Computer Vision- DSE312
ANSHUMAN DANGWAL 21044, SRUTANIK BHADURI 21275

Anonymous authors
Paper under double-blind review

ABSTRACT

Road object detection is a vital aspect of computer vision and autonomous driving, involving the use of algorithms and deep learning models to identify and classify objects on roadways, such as vehicles, pedestrians, and obstacles. This technology, often leveraging convolutional neural networks, is trained on diverse datasets to ensure accurate detection in various scenarios. Challenges include handling occlusions and complex interactions between objects. Road object detection enhances safety and navigation in applications ranging from advanced driver assistance systems to autonomous vehicles, with ongoing research focused on improving accuracy and adaptability to real-world driving conditions.

1 INTRODUCTION

The project is motivated by the need for precise road object detection in autonomous driving and advanced driver assistance systems. Leveraging the capabilities of 3D-RetinaNet, the study addresses the challenges presented by the Oxford Road Dataset, a widely recognized benchmark for evaluating road scene understanding algorithms.

2 DATASET

The ROAD dataset is specially designed from the per- spective of self-driving cars, and thus includes actions per- formed not just by humans but by all road agents in spe- cific locations, to form road events (REs). REs are anno- tated by drawing a bounding box around each active roadagent present in the scene, and linking these bounding boxes over time to form 'tubes'. As explained, to this purpose three different types of labels are introduced, namely: (i) the category of road agent involved (e.g. Pedestrian, Car, Bus, Cyclist); (ii) the type of action being performed by the agent (e.g. Moving away, Moving towards, Crossing and so on), and (iii) the location of the road user relative the au- tonomous vehicle perceiving the scene (e.g. In vehicle lane, On right pavement, In incoming lane).
ROAD is composed by 22 videos from the publicly avail- able Oxford RobotCar Dataset [56] (OxRD) released in 2017 by the Oxford Robotics Institute2 , covering diverse road scenes under various weather conditions.

3 METHODOLOGY

3.1 DATA PREPROCESSING

The Oxford Road Dataset, consisting of LiDAR point clouds and corresponding image frames, serves as the primary data source. Preprocessing involves careful handling of the dataset to ensure representation across diverse driving scenarios. First, the original sets of video frames were downloadedand demosaiced, in order to convert them to red, green, and blue (RGB) image sequences. Then, they were encoded into proper video sequences using ffmpeg3 at the rate of 12 frames per

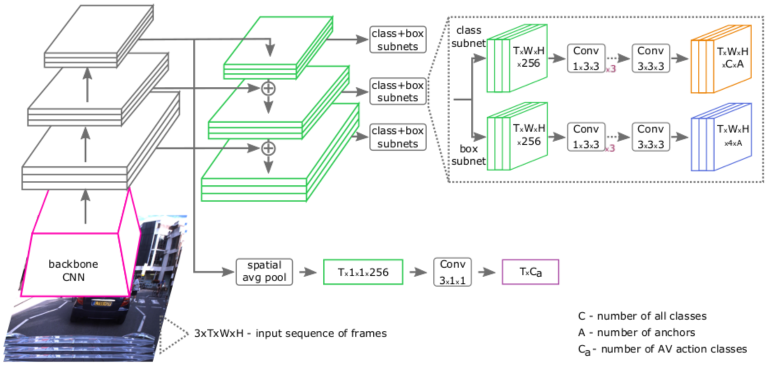

Figure 1: 3D-RetinaNet architecture

second (fps).First, the original sets of video frames were downloaded and demosaiced, in order to convert them to red, green, and blue (RGB) image sequences. Then, they were encoded into proper video sequences using ffmpeg3 at the rate of 12 frames per second (fps). The annotations information is stored in a json file having the same name as the video. The file structure contains the bounding boxes' coordinates and the associated labels per frame.

3.1.1 Quantitative Summary

Overall, 122K frames extracted from 22 videos were la- belled, in terms of both AV own actions (attached to the en- tire frame) and bounding boxes with attached one or more labels of each of the three types: agent, action, location. In total, ROAD includes 560K bounding boxes with 1.7M instances of individual labels.

3.2 3D-RetinaNet

3D-RetinaNet (4) is an innovative architecture designed to enhance object detection by extending the traditional RetinaNet into the three-dimensional domain, particularly well-suited for applications such as autonomous driving. The architecture employs a feature pyramid network (FPN) to extract hierarchical features from the input data, ensuring that the model can effectively detect objects of different sizes.

As in classical FPNs (1), the initial block of 3DRetinaNet consists of a backbone network outputting a series of forward feature pyramid maps, and of lateral layers producing the final feature pyramid composed by T feature maps. The second block is composed by two sub-networks which process these features maps to produce both bounding boxes (4 coordinates) and C classification scores for each anchor location (over A possible locations).

As in FPN, we adopt ResNet50 (2) as the backbone network. The 3D aspect is incorporated by extending these features into the temporal dimension, creating a fusion of spatial and temporal information. This comprehensive integration of spatial and temporal features equips 3D-RetinaNet to excel in scenarios where understanding the evolving nature of objects over time is critical, making it particularly valuable for applications like road object detection in autonomous vehicles.

4 Results

The results are evaluated in terms of both frame-level bounding box detection and of tube detection. In the first case, the evaluation measure of choice is frame mean average precision (f-mAP). We set the Intersection over Union (IoU) detection threshold to 0.5 (signifying a 50% overlap between predicted and true bounding box).

The I3D network achieves 75.2 fmap score while 2D achieves 65.2

3D-RetinaNet / 2D (ours)*	65.2
3D-RetinaNet / I3D (ours)	**75.2**

Figure 2: Results

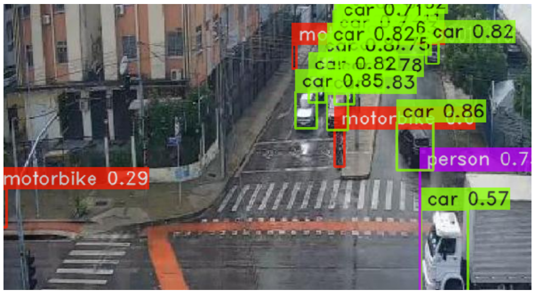

Figure 3: YOLO NASA results sample

5 YOLO NAS

For each detection task except agentness (which amounts to object detection) the performance is quite lower than the 75.2% achieved by our 3D baseline network. This is due to the numerous nuisance factors present in ROAD, such as significant camera motion, weather conditions, etc and lack of computation power.
Therefore we tried road object detection using YOLO NAS(3)(You Only Look One - Neural Architecture Search) in a much smaller dataset.
YOLO NAS combines the strengths of the YOLO (You Only Look Once) object detection framework with neural architecture search, enabling the automatic discovery of network architectures tailored for precise and fast detection of road objects. The study evaluates the performance of YOLO NAS on road scenes, aiming to contribute to the advancement of efficient and accurate road object detection systems.
The dataset contains various images of traffic. Images mostly taken from Turkey The results we got were far better than the results on ROAD dataset. The confidence score on true positives was ranging from 0.82-0.94.

References

[1] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[2] Sheldon Mascarenhas and Mukul Agarwal. A comparison between vgg16, vgg19 and resnet50 architecture frameworks for image classification. In *2021 International conference on disruptive technologies for multi-disciplinary research and applications (CENTCON)*, volume 1, pages 96–99. IEEE, 2021.

[3] Mantu Naresh Sharma. Image and video segmentation using yolo-nas and segment anything model (sam): Machine learning.

[4] Gurkirt Singh, Stephen Akrigg, Manuele Di Maio, Valentina Fontana, Reza Javanmard Alitappeh, Salman Khan, Suman Saha, Kossar Jeddisaravi, Farzad Yousefi, Jacob Culley, et al. Road: The road event awareness dataset for autonomous driving. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):1036–1054, 2022.