# Cameras vs LiDAR Using Deep Learning

Fernando Rojas Ramos and Ivo Pineda Torres

August 14, 2021

# Cameras vs LiDAR using Deep Learning

1st Fernando Rojas Ramos
*Facultad de Ciencias de la Computación*
*Benemérita Universidad Autónoma de Puebla*
Puebla, México
toshk15@hotmail.com

2nd Ivo H. Pineda Torres
*Facultad de Ciencias de la Computación*
*Benemérita Universidad Autónoma de Puebla*
Puebla, México
ivopinedatorres@gmail.com

***Abstract*—Research centers and companies dedicated to the development of autonomous vehicles are opting for two trends: using only cameras as a vision system or using LiDAR sensors plus cameras. Tesla has a fully camera vision system and Waymo has a cameras and LiDAR system. The difference of the LiDAR sensor could prevent accidents and save human lives in the future. Therefore, the main contribution of this work is the design of a methodology based on the comparison of detection efficiency for vision devices (Cameras and LiDAR). Applying measurement parameters provided by Neural Networks and models evaluation metrics in machine learning; it has to be concluded if its necessary to use LiDAR sensors in the development of autonomous cars.**

***Index Terms*—Autonomous Vehicles, Vision System, Neural Networks.**

## I. Introduction

Currently, the development of autonomous cars is of great interest to most automotive companies, including Tesla, Google, VW, Toyota, Ford and many others. The Waymo company provides a robot taxi service in the city of Phoenix in the United States; this is a limited service for this region and with their cars being equipped with a vision system based on cameras, radar and LiDAR sensors. Their aim is to autonomously keep their vehicles centered on the lane and to reach the passenger's destination.

Moreover, the Tesla company is also in the fight to achieve level 5 autonomy for its cars, which include advanced hardware that currently provides Autopilot functions, and full autonomous driving functionalities in the near future through updates. While their software is designed to improve functionality as time goes on, it has, however run, into many difficulties with a considerable number of car accidents. Put on doubt its vision system based only on cameras and radar-type sensors, Elon Musk CEO of Tesla is against using a LiDAR sensor.

Tesla has been heavily relying on vision and going against LiDAR sensors. Simultaneously, all the other companies use LiDAR and seem to dismiss other options.

The most apparent reason for Tesla has taken a different route is the cost. The cost of placing a single LiDAR device on a car is somewhere around $10,000. Google with its Waymo project has been able to slightly decrease the number by introducing mass production. However, the cost is still rather significant.

Car accidents are the eighth leading cause of death worldwide with 95% being caused by human error; the expectation is that the automation of transport represents a significant reduction in the number of occurrences and mainly of victims.

Nowadays, many current image-based object detectors using convolutional neural networks exhibit excellent performance on existing datasets such as KITTI [1]. This dataset will be used to carry out the study with images and real data from LiDAR sensors.

The main contributions of this work is the development of a result evaluation-based methodology to compare the detection efficiency of both devices (Cameras and LiDAR). who are based on the measurement parameters provided by neural networks and model evaluation metrics in machine learning.

This paper is organized as follows: Section 2 gives a brief analysis of the most related papers. Section 3 proposes methodology to analyze the performance of the built detectors. Finally, the conclusions and future work in Section 4.

## II. Related Work

There are a variety of studies aimed at the autonomy of automobiles and, in particular, at devices that function as a means of vision - in this case LiDAR sensors and Cameras. The studies related to this project are focused on examining the detection of objects separately, that is, the results are based on the detection of objects using only cameras. [2]–[5] or LiDAR sensors. [6]–[9]. In this project, the two trends are covered to carry out a methodology for comparing the results. This serves to determine which device provides the best detection.

Manuel Herzog and Klaus Dietmayer [10], propose in their work the detection of objects with LiDAR sensors in driverless cars and by applying the strategy of training a model using data from different types of LiDAR sensors. In comparison to the author Haris, M. [11]: this work proposes the detection of small and medium obstacles that were left on the road intentionally or unintentionally, which can pose a danger for both autonomous and human driving situations. They use Random Markov Field (MRF) models by merging three potentials (gradient potential, curvature prior potential, and depth variance potential) to segment obstacles and non-obstacles in the hazardous environment. Finally, they use obstacle detection to predict the steering wheel angle of the autonomous car using images from cameras.

Identify applicable funding agency here. If none, delete this.

Di Feng [12]: he tries to summarize systematically the methodologies in his article and to discuss the challenges for deep multimodal object detection and semantic segmentation in autonomous driving. To this end, they first provided an overview of sensors embedded in test vehicles, open datasets, and basic information for object detection and semantic segmentation in autonomous driving research.

The technological trend in using sensor fusion for the development of autonomous cars is of great interest and study for the coming years because it involves the integration of all data from radars, LiDAR, cameras, gps, etc.

Guan [13]: propose in his work the fusion of sensors to achieve better performance in the detection of objects. This consists of a multi-adaptive and high precision completion method which improves the adaptability to the detection environment and performs preliminary fusion of data from two sensors (Camera and LiDAR). The system realized fast and accurate object detection through the real-time object detection model called YOLOv3 and applied a proposed decision-level fusion strategy. The methodology not only gets higher detection precision during daytime driving but also obtains the distance between the front vehicle and the detecting vehicle.

Wangs' work [14] is based on the use of stereo cameras for the detection of objects, these images have the advantage of showing more information than the images of monocular cameras; they use depth in image scenes to locate how far away the detected object is as with LiDAR sensors. With these images they are able to simulate the operation of the LiDAR sensors. However, they also prove that the fusion of the images from the cameras plus the point cloud of the LiDARs obtain better detection results. The article proposes as future work to compare the processing times of the data from both detection devices.

## III. Proposed Methodology

In order to solve the problem of object detection using deep learning, the methodology shown in Fig. 1 is proposed. A convolutional neural network called U-net is used for the segmentation of labeled objects in the dataset KITTI for their subsequent detection in a bounding box and is also used the convolutional neural network called YOLOv5 recently published to compare the detection results, which is the objective of this work.

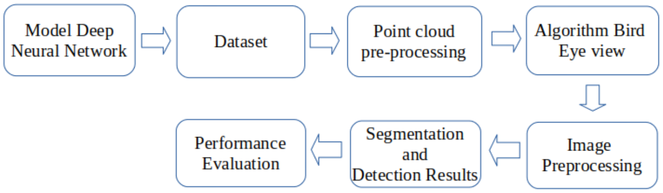
Model Deep Neural Network
Dataset
Point cloud pre-processing
Algorithm Bird Eye view
Image Preprocessing
Segmentation and Detection Results
Performance Evaluation

Fig. 1.

### A. Models Deep Neural Networks U-net and YOLOv5

The U-Net was developed by Olaf Ronneberger [15] for the segmentation of biomedical images. The architecture contains two paths. The first path is the shrink path (also called the encoder) that is used to capture the context in the image. The encoder is just a traditional stack of maximum grouping and convolutional layers. The second path is the symmetric expansion path (also called a decoder) that is used to allow precise localization by transposed convolutions. So it is a fully convolutional end-to-end network (FCN), that is, it only contains convolutional layers and does not contain any dense layers due to which it can accept images of any size.

The YOLOv5 [16] got released by Glenn Jocher(Founder and CEO of Utralytics) and has been chosen for the task of object detection and for its incredible characteristics: the speed and excellent precision in the detection of objects. These are the following versions of YOLO with enhancements for the detection of small objects: YOLOv1, YOLOv2, YOLOv3, YOLOv4 and the latest recent YOLOv5 release. This model has the capacity to process 140 frames per second with the disadvantage of the first version of not accurately detecting small objects, this problem has been improved with the development of the existing versions. YOLO has an average precision (mAP) value of 57.9% on the COCO dataset, which is significantly higher than an SSD type network and a RetinaNet, being 4 times faster than them, 100 times faster than a Fast R-CNN.

### B. KITTI Dataset

KITTI [1] is a dataset available to carry out the study with real data of a prototype autonomous vehicle. The vehicle is equipped with four cameras: 1 stereo pair of color cameras and 1 stereo pair of grayscale cameras. The color and grayscale cameras are mounted close to each other ( 6 cm) the baseline of both stereo decks is approximately 54 cm. This configuration allows to obtain information in both color and grayscale from the left and right camera. While color cameras (obviously) come with color information, grayscale camera images have higher contrast and slightly less noise.

All cameras are clocked at approximately 10 Hz relative to the Velodyne laser scanner. The trigger is mounted in such a way that the images from the camera roughly coincide with Velodyne lasers facing forward (in the driving direction).

All camera images are provided as lossless compressed and rectified png sequences. Native image resolution is 1382x512 pixels and slightly lower after rectification.

The classes available in the KITTI dataset are 8 and the number of instances are the following, Car = 28614, cyclist = 612, Misc = 959, Pedestrian = 4448, Person sitting = 220, Tram = 504, Truck = 1079, Van = 2900.

### C. Point cloud preprocessing

Point clouds are a collection of points that represent a 3D shape or feature. Each point has its own set of X, Y and Z coordinates and in some cases additional attributes.

Nowadays, object detection systems can be divided into two main categories . The first ones are the geometric based, which retrieve the obstacles using geometric and morphological operations on the 3D points. The seconds are the deep learning based, which process the 3D points, or an elaboration of the 3D point cloud, with deep learning techniques to retrieve a set of obstacles.

This work is focused on the second approximation: projection based methods implement a single or multi-view projection of a 3D point cloud, resulting in a 2D grid, which is then processed to find object clusters with the desired confidence. Afterwards, this grid is processed by is then processed by a 2D Convolution neural network.

The first thing to do is bring the 3D point cloud of the LiDAR sensors to a 2D voxel type image, with the alignment of the points in space and time according to the calibration algorithm of the cameras with the LiDAR sensor. The calibration algorithm calculates the camera matrix using the extrinsic and intrinsic parameters. The extrinsic parameters represent a rigid transformation from 3D world coordinate system to the 3D cameras coordinate system. The intrinsic parameters represent a projective transformation from the 3D cameras coordinates into the 2D image coordinates. As shown in our diagram see Fig. 2, we have the image of the camera and the point cloud of the LiDAR sensors; the objective is the synchronization of the points with the pixels of the image so that the 2D surface of our new image coincides with the projection of the 3D point cloud. Using the camera's calibration parameters, we can carry out the transformation of the point cloud to a 2D voxel image as the camera does.

### *D. Algorithm Bird eye view*

The perspective transform that interests us is a birds-eye view transform [17] that enables us to view a lane from above. Aside from creating a birds eye view representation of an image, a perspective transform can also be used for all kinds of different view points.

Channel feature extraction [18] the 3D point-cloud provided by the LiDAR is projected in a BEV image with predetermined width, height and grid cell size. To avoid loss of information during the projection of 3D point-cloud into a 2D image, 6 additional channels are stacked together the new pattern to recover information about the peak and the medium values of height, intensity and distribution of the collapsed points for each cell. Moreover, binary information concerning the effective occupancy of each grid are included.

### *E. Image Preprocessing*

This phase involves the preprocessing of digital images to provide them with different attributes with data augmentation [19], [20]. In the real world scenario, it's possible that the existing datasets are taken under a limited conditions. This is why data augmentation is applied to simulate different conditions and generate a random variety images.

Popular augmentation techniques: lightening the image to increase their clarity and avoiding dark images; decreasing the contrast to reduce the total color range of the image; applying a Gaussian filter to eliminate noise and details of the texture; flipping images horizontally and vertically; rotating it at right angles will preserve the image size; scaled outward or inward; the method of resizing the section is popularly known as random cropping; translation just involves moving the image along the X or Y direction (or both).

### *F. Segmentation and Detection Results*

According to the design of the methodology, the necessary phases have been implemented to obtain the results of segmentation and detection in images of the cameras Fig. 3 and images from the data of the LiDAR sensors Fig. 4. In this phase the experiments are carried out successfully with the Car class which has the highest number of labels and, therefore, presents balance with respect to the rest of the classes available in the KITTI dataset.

The processes described in the methodology have been run on a computer with a GeFORCE GTX 1050 graphics card and 24 GB of RAM necessary for the process of generating the images with bird eye view from the point cloud of the LiDAR sensors; this takes extra time compared to the images from cameras. This phase of the methodology is of primary interest-working with other types of sensor data to train a neural network. The results present four images per row as follows in this case for the neural network U-net: The first image represents the original test image; the second image shows the total targets to be predicted; the third image shows the prediction of the object's segmentation by the U-net neural network; ultimately, the fourth image with a bounding box represents the detection of the segmentation predicted by the neural network with a threshold greater than or equal to 0.5.

The following results are those obtained with the convolutional neural network YOLOv5. The object detection is performed directly on the test images with a bounding box and a label with the name of the class. During training, the YOLOv5 training pipeline creates batches of training data with augmentations. We can visualize the training data ground truth as well as the augmented training data. The first figure 5 shows the results obtained with the images of the cameras and the second figure 6 shows the results with the LiDAR sensor.

### *G. Performance Evaluation*

Here are three metrics used as reference; these are precision, recall, and measure F1, which assess the balance between precision and recall.

$$precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{1}$$

$$recall = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{2}$$

$$F1\,measure = 2 \times \frac{precision \times recall}{precision + recall} \tag{3}$$

Based on the confusion matrix, the evaluation metrics corresponding to the predictions of the two deep neural networks are calculated.

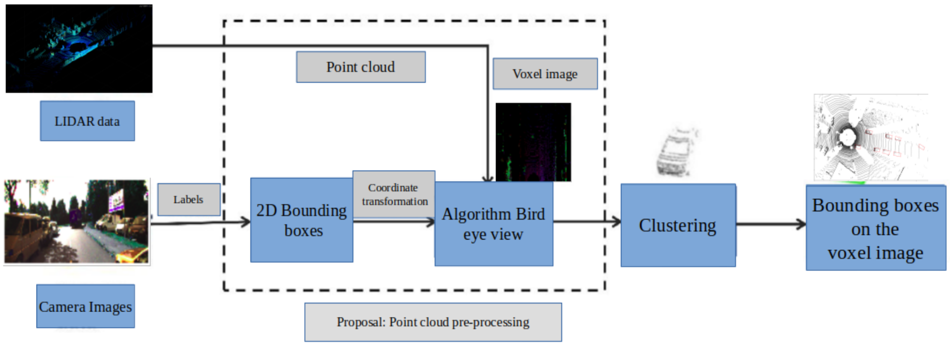

Fig. 2. The coordinate conversion between the image and LiDAR

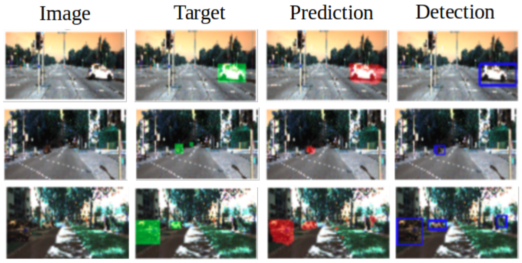

Fig. 3. Segmentation and detection of the ”Car” class with Camera.

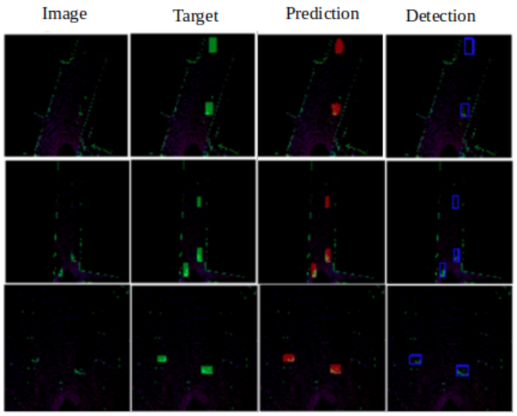

Fig. 4. Segmentation and detection of the ”Car” class with LiDAR.

The following tables I and II, shows the evaluation metrics with respect to the predictions made by deep neural networks U-net and YOLOv5. It is important to mention that at this point the only class that is being evaluated is ”Car”, as described it is the class with the largest number in instances, unlike the other classes available in the dataset.

## IV. Conclusion and Future Work

In this work, it was possible to integrate advanced computer vision algorithms using deep neural networks, whose results were positive in recent studies conducted in autonomous prototype vehicles.

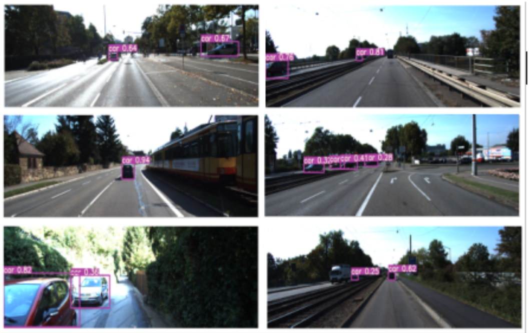

Fig. 5. Detection of the ”Car” class with Camera.

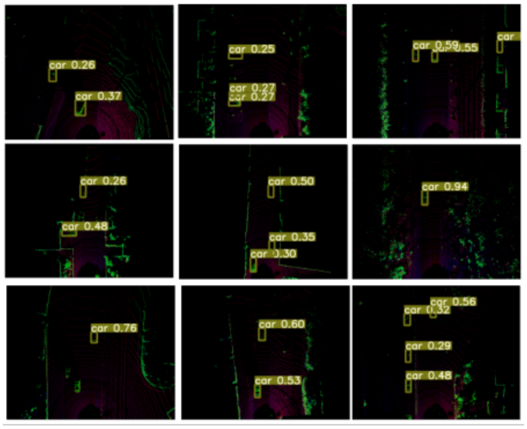

Fig. 6. Detection of the ”Car” class with LiDAR.

Data from the LiDAR sensors were analyzed which provides us with a 360 degree view with a single sensor compared to cameras. The LiDAR sensor provides good results in the segmentation and detection of objects; therefore, the proposal is to use it in conjunction with the cameras to develop a sophisticated and complete vision system. This means, that if at some point the cameras do not detect an object while driving, the LiDAR sensor is supported as security.

The depth estimation and the creation of maps of the

TABLE I
TEST SET 100 IMAGES AND TOTAL 261 INSTANCES.

| Model | TP | FN | TN | FP | Presicion | Recall | F1 measure |
|---|---|---|---|---|---|---|---|
| **YOLOv5 Camera** | 176 | 85 | 0 | 0 | 100.00% | 67.00% | 80.00% |
| **YOLOv5 LiDAR** | 144 | 48 | 0 | 0 | 100.00% | 75% | 85.71% |
| **U-net Camera** | 123 | 45 | 53 | 40 | 75.46% | 73.21% | 74.31% |
| **U-net LiDAR** | 132 | 42 | 55 | 38 | 77.64% | 75.80% | 76.70% |

TABLE II
TEST SET 200 IMAGES AND TOTAL 482 INSTANCES.

| Model | TP | FN | TN | FP | Presicion | Recall | F1 measure |
|---|---|---|---|---|---|---|---|
| **YOLOv5 Camera** | 336 | 142 | 0 | 4 | 98.8% | 70.00% | 81.90% |
| **YOLOv5 LiDAR** | 264 | 84 | 0 | 2 | 99.2% | 75.80% | 85.90% |
| **U-net Camera** | 233 | 88 | 99 | 82 | 73.96% | 72.58% | 73.26% |
| **U-net LiDAR** | 254 | 83 | 98 | 80 | 76.04% | 75.37 % | 75.70% |

T

LiDAR sensors help to detect and locate the objects within the environment. These characteristic could allow to avoid accidents such as those that have occurred with the Tesla brand cars.

In this work, we conclude the efficiency of the LiDAR sensor for its instrumentation in autonomous cars.

As future work, it is proposed to analyze the possibility of equipping a car with a LiDAR sensor and cameras to generate its own dataset on the roads of Mexico.

## Acknowledgment

We gratefully acknowledge to Conacyt with the financial support to carry out this research.

## References

[1] Geiger, A., Lenz, P., and Urtasun, R. Are we ready for autonomous driving? the kitti vision benchmark suite. In 2012 IEEE Conference on Computer Vision and Pattern Recognition (pp. 3354-3361). IEEE.

[2] Ramesh, B., Ussa, A., Della Vedova, L., Yang, H., and Orchard, G. Low-power dynamic object detection and classification with freely moving event cameras. Frontiers in neuroscience, 14, 135, 2020.

[3] Meyer, M., and Kuschk, G. Deep learning based 3d object detection for automotive radar and camera. In 2019 16th European Radar Conference (EuRAD) (pp. 133-136). IEEE.

[4] Nobis, F., Geisslinger, M., Weber, M., Betz, J., and Lienkamp, M. A deep learning-based radar and camera sensor fusion architecture for object detection. In 2019 Sensor Data Fusion: Trends, Solutions, Applications (SDF) (pp. 1-7). IEEE.

[5] Fuentes-Jimenez, D., Martin-Lopez, R., Losada-Gutierrez, C., Casillas-Perez, D., Macias-Guarasa, J., Luna, C. A., and Pizarro, D. DPDnet: A robust people detector using deep learning with an overhead depth camera. Expert Systems with Applications, 146, 113168, 2020.

[6] Chen, C., Fragonara, L. Z., and Tsourdos, A. RoIFusion: 3D Object Detection From LiDAR and Vision. IEEE Access, 9, 51710-51721, 2020.

[7] Meyer, G. P., Laddha, A., Kee, E., Vallespi-Gonzalez, C., and Wellington, C. K. Lasernet: An efficient probabilistic 3d object detector for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12677-12686), 2019.

[8] Lang, A. H., Vora, S., Caesar, H., Zhou, L., Yang, J., and Beijbom, O. Pointpillars: Fast encoders for object detection from point clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12697-12705), 2019.

[9] Sahba, R., Sahba, A., Jamshidi, M., and Rad, P. 3D Object Detection Based on LiDAR Data. In 2019 IEEE 10th Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON) (pp. 0511-0514). IEEE.

[10] Herzog, M., Dietmayer, K. Training a fast object detector for lidar range images using labeled data from sensors with higher resolution. In 2019 IEEE Intelligent Transportation Systems Conference (ITSC) (pp. 2707-2713). IEEE.

[11] Haris, M., and Hou, J. (2020). Obstacle detection and safely navigate the autonomous vehicle from unexpected obstacles on the driving lane. Sensors, 20(17), 4719.

[12] Feng, D., Haase-Schuetz, C., Rosenbaum, L., Hertlein, H. and Dietmayer, K. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. IEEE Transactions on Intelligent Transportation Systems, 2020.

[13] Guan, L., Chen, Y., Wang, G., and Lei, X. Real-Time Vehicle Detection Framework Based on the Fusion of LiDAR and Camera. Electronics, 9(3), 451, 2020.

[14] Wang, Y., Chao, W. L., and Garg, D. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8445-8453), 2019.

[15] Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham, 2015.

[16] https://github.com/ultralytics/yolov5

[17] Yang, G., Mentasti, S., Bersani, M., Wang, Y., Braghin, F., and Cheli, F. LiDAR point-cloud processing based on projection methods: a comparison. In 2020 AEIT International Conference of Electrical and Electronic Technologies for Automotive (AEIT AUTOMOTIVE) (pp. 1-6). IEEE.

[18] Lee, K. H., Kliemann, M., Gaidon, A., Li, J., Fang, C., Pillai, S., and Burgard, W. PillarFlow: End-to-end Birds-eye-view Flow Estimation for Autonomous Driving. arXiv e-prints 2020, arXiv-2008.

[19] Tian, Y., Pei, K., Jana, S., and Ray, B. Deeptest: Automated testing of deep-neural-network-driven autonomous cars. In Proceedings of the 40th international conference on software engineering (pp. 303-314), 2018.

[20] Du, S., Guo, H., and Simpson, A. Self-driving car steering angle prediction based on image recognition, 2019. arXiv preprint arXiv:1912.05440.