HAL
open science

A new Evaluation Approach for Deep Learning-based Monocular Depth Estimation Methods

Antoine Mauri, Redouane Khemmar, Rémi Boutteau, Benoit Decoux, Jean Yves Ertaud, Madjid Haddad

To cite this version:

Antoine Mauri, Redouane Khemmar, Rémi Boutteau, Benoit Decoux, Jean Yves Ertaud, et al.. A new Evaluation Approach for Deep Learning-based Monocular Depth Estimation Methods. The 23rd IEEE International Conference on Intelligent Transportation Systems, Sep 2020, Rhodes (virtual conference), Greece. hal-02978149

HAL Id: hal-02978149

https://hal.science/hal-02978149v1

Submitted on 26 Oct 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A new Evaluation Approach for Deep Learning-based Monocular Depth Estimation Methods

Antoine Mauri [1,2], Redouane Khemmar [1], Remi Boutteau [1], Benoit Decoux [1], Jean-Yves Ertaud [1] and Madjid Haddad[2]

***Abstract*— In smart mobility based road navigation, object detection, depth estimation and tracking are very important tasks for improvement of the environment perception quality. In the recent years, a surge of deep-learning based depth estimation methods for monocular cameras has lead to significant progress in this field. In this paper, we propose an evaluation of state-of-the-art depth estimation algorithms based on single single input on both the KITTI dataset and the recently published NUScenes dataset. The models evaluated in this paper include an unsupervised method (Monodepth2) and a supervised method (BTS). Our work lies in the elaboration of novel depth evaluation protocols, object depth evaluation and distance ranges evaluation. We validated our new protocols on both KITTI and NUScenes datasets, allowing us to get a more comprehensive evaluation for depth estimation, especially for applications in scene understanding for both road and rail environment.**

I. INTRODUCTION

Accurate depth estimation is necessary for the perception of the environment in front of the vehicle and can significantly increase safety by estimating the distance of pedestrians and vehicles. Technology can also improve the competitiveness of road transports, as shown in [1], [2], but this field still has many important challenges before being completely operational. Distance measurement from objects can be based on many kind of sensors: ultrasonic, laser [3] or time-of-flight camera [4]. But those solutions are still costly. In this work, we investigate the use of camera(s) for this task. The most common method to estimate distance with cameras is to use a pair of stereoscopic cameras with matching algorithms. But recently, vision algorithms based on Convolutional Neural Networks (CNN) have shown state-of-the-art performance in depth estimation with a single camera. Such methods has the advantages to require a relatively low-cost and easy to integrate sensor. But those methods still lack in precise evaluation. In this context, several works related to environment perception have already been carried out, such as the tracking of a person [5], the detection and tracking of objects for the road smart mobility [6][7].

While the depth estimation algorithms tested in this paper offer comprehensive evaluation results, they only provide an evaluation of the overall performance of a method. It lacks information of how well objects have their distance predicted and what is the depth precision at longer ranges. These informations are vital for applications in road environment especially for autonaumous driving. That is what motivated our contribution of an evaluation protocol better suited for road environments as well as a comparative evaluation using our new protocols of state-of-the-art methods on two large datasets on road environments: KITTI [8] and NuScene [9].

The main contribution of our work is to offer a new evaluation protocol for single image depth estimation algorithms adapted to road and rail environment for autonomous vehicles as well as an evaluation of state-of-the-art methods.

The remainder of this paper is organized as follows: In section II we review the state-of-the-art algorithms which are evaluated in this paper, the datasets which are used, and present the new proposed protocols. In section III, we describe in more details the new class-specific metric which is proposed, the methodology of evaluation and some specificities of learning on the NuScene dataset. Experimental results are presented in section IV. Finally, conclusion and future directions are drawn in section V.

II. RELATED WORK

A. Monocular Depth Estimation Methods

For learning purposes, if the datasets used contains the distance from object as ground-truth information, this information can be used to supervise learning of a neural network with a regression output layer. Most of the CNN models for depth estimation have an encoder-decoder structure, similar to the one that is used for the application of semantic segmentation of images [10] [11], in which the output of the network has the same size as the inputs. One of the main problems encountered by those models is to get full resolution at the output of the network, due to the bottleneck which exists at the junction between the encoder and decoder parts.

In Multi-Scale Local Planar Guidance model (called BTS for Big-To-Small in the following, the name given by the authors in their paper) [12], layers located at multiple stages of the decoding phase are used and their outputs are combined to predict depth at full resolution. Another CNN in this category is DenseDepth [13], in which the encoder part is based on a pre-trained DenseNet [14]. The whole network is then trained with NYU Depth v2 [15] and KITTY datasets. A specific loss function penalizing high-frequency distorsion allows for more faithfull depth reconstruction at object boundaries.

Another way to get supervision information is to use aligned pairs of stereo images during the learning phase, and then infer depth maps on monocular images. This approach

[1]Normandie Univ, UNIROUEN, ESIGELEC, IRSEEM, 76000 Rouen, France.
[2]SEGULA Engineering, 19 rue d'Arras, 92000 Nanterre, France.

is called self-supervised, as those models don't need dataset with ground-truth (like LiDAR measurement or disparity maps) for learning. In Monodepth [16], the problem of depth estimation is casted into an image reconstruction one. Specifically, the CNN is given as input one of the images of the stereo pair, and gives as output an estimate of the disparities. The second image of the stereo pair is then synthesized from this estimate. The difference between the synthesized image and the real one is then used in the loss function [17][18]. In MonoResMatch model [18], authors also use stereo images with an end-to-end learning to get the depth information. They do not use ground truth present in the datasets, but use a stereo-matching algorithm (Semi-Global Matching) to internally generate this information. In Monodepth2 [19], authors present a set of improvements over the first version of their algorithm [16]. The model is based on three processes which collaborate together in order to improve depth estimation: the model can be trained with monocular data, stereo data, or both. Specifically, the 3 processes are made of: (1) a minimum reprojection loss calculated for each pixel, (2) an auto-masking loss to ignore confusing and (3) a full resolution multi-sampling method for reducing visual artifacts. The effectiveness of the model is demonstrated quantitatively and qualitatively on the KITTI benchmark.

Other methods use monocular images for both learning and inference. Specifically, sequences of images are used for learning, and at test time depth is estimated on single images. In [20], a CNN with two modules sharing the first few convolutional layers, is used to jointly give estimates for depth and pose. In [21], geometric structure of objects and of the scene are introduced in the learning process, by using separate object motion estimators, ego-motion estimator and depth estimator, making the model well adapted to highly dynamic environments.

By using sequences of stereo images, it is also possible to learn depth and odometry at the same time, as both spatial and temporal photometric errors are available [22].

B. Datasets for Depth Prediction in Outdoor Environment

In smart mobility based road and railway navigation, object detection, depth estimation and tracking are very important tasks for improvement of the environment perception quality. A deep learning method for object detection and depth estimation requires more and more training and evaluation datasets containing heterogeneous data like images, videos, ranges, etc. This is why it is very important to identify a good and high accuracy dataset for our real time object detection and depth estimation for road and rail applications.

In KITTI [8], the authors present one of the highly used dataset in road environment for mobile robotics and autonomous driving research. KITTI is a calibrated, synchoronized, and timestamped autonomous driving dataset which was captured with a wide range of scenarios. The KITTI dataset was collected by using a VW Station vehicle instrumented by different kinds of sensors such as: color and grayscale stereo cameras, a velodyne 3D laser scanner and a high precision GPS/IMU navigation system. The platform contains real-world traffic situations with both static and dynamic objects (object labels are presented in the form of 3D tracklets). The dataset provides online benchmarks for different tasks such as: stereo, optical flow, and object detection.

NuTonomy scenes (NUScenes [9]) is a multimodal dataset for autonomous driving. NUScenes contains different types of sensor surch as: 6 cameras, 5 radars and 1 LiDAR. It is fully annotated and comprises 1000 scenes ($20s$ long for each), 3D bounding boxes for 23 classes and 8 attributes. It has 100x as many images than the pioneering KITTI dataset. It also contains careful dataset analysis as well as baselines for LiDAR and image based detection and tracking [9].

III. Evaluation of Deep learning-based Monocular Depth Estimation Methods

A. Error Metrics used in Depth Evaluation

Before presenting our contributions for depth evaluation, we define below the depth error metrics that are used in the literature and in our work. Let p be the depth prediction of a pixel in the image, g its ground truth and N the total number of depth pixels in the image.

Relative Error: The equation for the Relative Error (RE) is detailed in Equation (1).

$$RE = \frac{1}{N}\sum_i\sum_j \frac{|g_{i,j} - p_{i,j}|}{g_{i,j}} \tag{1}$$

Squared Relative Error: The equation for the Squared Relative Error (SRE) is detailed in Equation (2).

$$SRE = \frac{1}{N}\sum_i\sum_j \frac{|g_{i,j} - p_{i,j}|^2}{g_{i,j}} \tag{2}$$

Root Mean Squared Error: The Root Mean Squared Error (RMSE) details can be found in Equation (3).

$$RMSE = \sqrt{\frac{1}{N}\sum_i\sum_j (g_{i,j} - p_{i,j})^2} \tag{3}$$

Logarithmic Root Mean Squared Error: The equation for the Logarithmic Root Mean Squared Error (logRMSE) is detailed in Equation (4).

$$logRMSE = \sqrt{\frac{1}{N}\sum_i\sum_j (log(g_{i,j}) - log(p_{i,j}))^2} \tag{4}$$

Percentage of Bad Matching Pixels: The Percentage of Bad Matching Pixels (BMP) is detailed in Equation (5) where C is a threshold used for setting an error tolerance.

$$[a]_{k=[1..3]} = \frac{1}{N}\sum_i\sum_j max(\frac{g_{i,j}}{p_{i,j}}, \frac{p_{i,j}}{g_{i,j}}) < C^k \tag{5}$$

These metrics give a comprehensive statistical assessment of the performance of a method but we believe it can be further developed. One of our contributions lies in novel depth evaluation protocols that can be found below.

B. Object Depth Evaluation

While the current evaluation of depth estimations gives a comprehensive assessment of the overall performance of a given method, it is done on the global image and does not evaluate the object distance prediction. Object distance is a vital aspect for applications in autonomous driving and scene perception, this is why we designed a new depth evaluation protocol that allows us to compute the depth prediction error for relevant objects that are regularly encountered in road environments (person, car, truck, etc). Our evaluation protocol consists in 4 steps: (1) The predicted depth map is scaled using median scaling; (2) Object masks are generated using Mask-RCNN [23] (see Figure 1 for an example of network output); (3) The generated object masks are then used to segment the depth maps and the depth errors are computed for each mask in the image; (4) Finally the mean of the errors are computed for each class. This new evaluation protocol will allow a better understanding of how well a given method estimates the distance of objects present in road environment. It is especially useful for autonomous driving applications.

C. Depth Evaluation over Distance Ranges

The classic depth evaluation protocol also doesn't allow to evaluate a method's performance over longer ranges since the evaluation is done on the whole image. The performance of a method over longer ranges is an important parameter that needs to be taken into account for scene understanding purposes. Here we propose to follow the work of [24] where they described an evaluation protocol over distance ranges, but while they used this protocol for indoor scenes, we used it in a road environment where the distance ranges are more important. Our protocol is as followed: We scale the predicted depth map using median scaling. We then create the distance ranges of 10m up to a distance of 80m (i.e [0-10m], [10-20m], ..., [70-80m]). Each pixel is then assigned to a distance range according to the value of the depth ground truth. For each distance range, we compute the depth errors. This new protocol gives an assessment of how the depth estimation degrades over longer ranges.

IV. Results Analysis

In this section, we will present the results of our evaluation on both KITTI and NuScenes datasets of state-of-the-art supervised and self-supervised monocular depth estimation methods BTS and Monodepth2.

A. KITTI Dataset

For the evaluation on the KITTI dataset, we used the pretrained models provided by the authors both BTS and Monodepth2. The BTS model was trained with images from the Eigen training split [25] at a resolution of 704x352 and with a dense groundtruth. The pretrained weights of Monodeth2 were trained using the monocular training on Zhou's training split [20] at a resolution of 1024x320. The evaluation was performed on the eigen test split. The results of our evaluation of BTS and Monodepth2 can be found in Table I and II. The value of C for the Threshold error defined in Equation (5) has been set to 1.25.

B. NuScenes Dataset

For the evaluation on NuScenes, we had to train both methods on the training split of NuScenes. For BTS, we trained for 50 epoch with a batch size of 20 and a resolution of 192x192, we used the sparse data from the LiDAR supervision. Given that Monodetph2 is an unsupervised method, it relies on reprojection loss for the monocular training. If the training images have poor visibility, the training might not converge. That is why we selected the scenes from the training split where the visibility is good enough for the training to converge. We also used all the images from each scene and not just the images that were synced with the LiDAR in order to get a framerate high enough for the monocular training to work. We trained Monodepth2 for 20 epochs with a batch size of 12 and a training resolution of 446x224. The results of our evaluation of BTS and Monodepth2 can be found in Table III and IV. The value of C for the Threshold error defined in Equation (5) has been set to 1.25.

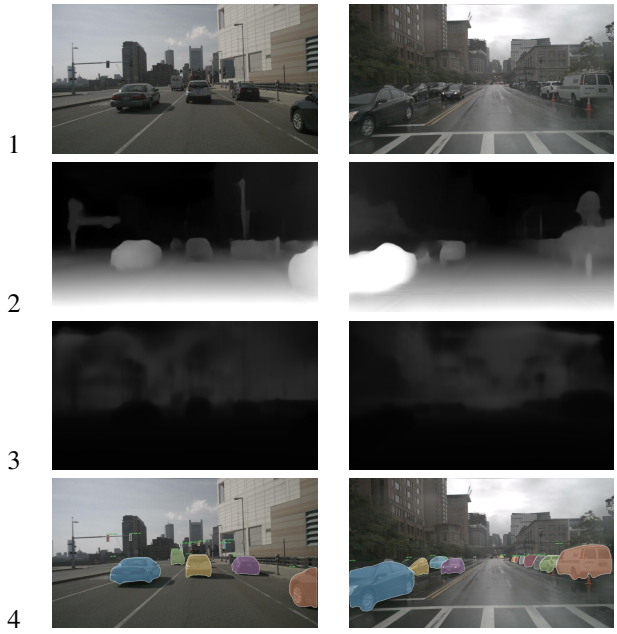

Fig. 1. Input data for our object distance evaluation protocol. (1) the input image fed to the depth prediction algorithm, (2) the disparity map, (3) the normalized depth map after median scaling and (4) the object masks from Mask-RCNN.

C. Experimental Results Analysis

Our results over the two datasets show that overall BTS yields better results than Monodepth2. Our evaluation over distance ranges also shows that both methods, as expected, tends to have a lower accuracy when the distance increases. Our object depth evaluation also shows that the depth estimation predictions errors for objects is significantly higher than the errors on the overall picture (see Figures 4 and 5), this can be explained by the large variety for each object class which

TABLE I

DEPTH EVALUATION OVER DISTANCE RANGES ON KITTI. THE ALGORITHMS EVALUATED ARE STATE-OF-THE-ART MONOCULAR DEPTH ESTIMATION METHODS: MONODEPTH2 (MD2) AND BTS. GLOBAL DEPTH ERRORS ARE SHOWN AS WELL AS DEPTH ERRORS FOR DISTANCE RANGES OF 10M AND UP TO 80M. BOTH SRE AND RMSE ARE EXPRESSED IN METERS.

	RE		SRE		RMSE		logRMSE		a_1		a_2		a_3	
Distance ranges	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS
$0-80m$	0.115	**0.060**	0.882	**0.249**	4.701	**2.798**	0.190	**0.096**	0.879	**0.955**	0.961	**0.993**	0.982	**0.998**
$0-10m$	0.102	**0.071**	0.503	**0.188**	1.489	**0.991**	0.141	**0.106**	0.929	**0.959**	0.979	**0.988**	0.99	**0.994**
$10-20m$	0.116	**0.088**	0.845	**0.395**	3.035	**2.198**	0.18	**0.149**	0.891	**0.924**	0.96	**0.971**	0.979	**0.985**
$20-30m$	0.168	**0.13**	1.866	**1.055**	6.208	**4.745**	0.261	**0.229**	0.773	**0.836**	0.916	**0.934**	0.957	**0.964**
$30-40m$	0.196	**0.16**	2.788	**1.945**	9.11	**7.476**	0.307	**0.279**	0.694	**0.764**	0.886	**0.906**	0.942	**0.947**
$40-50m$	0.209	**0.174**	3.504	**2.64**	11.682	**10.008**	0.318	**0.298**	0.641	**0.725**	0.865	**0.889**	**0.943**	0.941
$50-60m$	0.221	**0.19**	4.394	**3.739**	14.252	**12.852**	0.332	**0.326**	0.583	**0.675**	0.857	**0.868**	**0.927**	0.922
$60-70m$	0.212	**0.201**	4.657	**4.584**	15.855	**15.585**	**0.325**	0.334	0.609	**0.619**	0.854	**0.856**	**0.93**	0.923
$70-80m$	**0.181**	0.214	**4.34**	5.454	**15.8**	18.219	**0.284**	0.333	**0.652**	0.548	**0.873**	0.843	**0.945**	0.925

TABLE II

OBJECT DISTANCE EVALUATION ON KITTI. THE ALGORITHMS EVALUATED ARE STATE-OF-THE-ART MONOCULAR DEPTH ESTIMATION METHODS: MONODEPTH2 (MD2) AND BTS. DEPTH ERRORS WERE COMPUTED FOR THE OBJECT CLASSES WITH ENOUGH INSTANCE IN THE TEST SPLIT. BOTH SRE AND RMSE ARE EXPRESSED IN METERS.

	RE		SRE		RMSE		logRMSE		a_1		a_2		a_3	
Object class	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS
Person	0.314	**0.166**	5.721	**1.786**	8.43	**5.892**	0.326	**0.253**	0.601	**0.772**	0.829	**0.894**	0.92	**0.947**
Bicycle	0.131	**0.116**	0.517	**0.467**	2.81	**2.669**	0.172	**0.163**	0.829	**0.839**	**0.964**	0.962	0.993	**0.994**
Car	0.206	**0.137**	3.132	**1.491**	7.924	**6.052**	0.271	**0.223**	0.773	**0.838**	0.883	**0.922**	0.938	**0.955**
Truck	0.215	**0.122**	2.769	**0.826**	6.978	**4.523**	0.259	**0.177**	0.694	**0.854**	0.903	**0.969**	0.964	**0.985**

TABLE III

DEPTH EVALUATION OVER DISTANCE RANGES ON NUSCENES. THE ALGORITHMS EVALUATED ARE STATE-OF-THE-ART MONOCULAR DEPTH ESTIMATION METHODS: MONODEPTH2 (MD2) AND BTS. GLOBAL DEPTH ERRORS ARE SHOWN AS WELL AS DEPTH ERRORS FOR DISTANCE RANGES OF 10M AND UP TO 80M. BOTH SRE AND RMSE ARE EXPRESSED IN METERS.

	RE		SRE		RMSE		logRMSE		a_1		a_2		a_3	
Distance range	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS
$0-80m$	0.176	**0.147**	2.521	**1.184**	7.746	**5.849**	0.271	**0.214**	0.787	**0.817**	0.911	**0.94**	0.955	**0.977**
$0-10m$	**0.115**	0.116	0.982	**0.43**	1.561	**1.347**	**0.139**	0.151	**0.919**	0.915	0.972	**0.974**	0.986	**0.987**
$10-20m$	0.187	**0.153**	2.69	**0.966**	4.854	**3.452**	0.243	**0.197**	0.794	**0.81**	0.916	**0.945**	0.958	**0.984**
$20-30m$	0.242	**0.162**	3.488	**1.386**	8.316	**5.427**	0.32	**0.217**	0.643	**0.768**	0.859	**0.933**	0.932	**0.978**
$30-40m$	0.256	**0.183**	4.296	**2.01**	11.327	**7.922**	0.368	**0.255**	0.569	**0.677**	0.807	**0.909**	0.904	**0.969**
$40-50m$	0.264	**0.206**	5.14	**3.047**	14.167	**10.952**	0.404	**0.3**	0.518	**0.598**	0.76	**0.845**	0.888	**0.946**
$50-60m$	0.27	**0.225**	6.298	**4.441**	17.267	**14.398**	0.44	**0.346**	0.483	**0.556**	0.746	**0.789**	0.854	**0.904**
$60-70m$	0.28	**0.248**	8.024	**6.275**	20.725	**18.243**	0.475	**0.385**	0.48	**0.495**	0.692	**0.736**	0.817	**0.868**
$70-80m$	0.289	**0.284**	10.299	**8.802**	24.621	**23.16**	0.505	**0.434**	**0.463**	0.394	0.645	**0.677**	0.776	**0.834**

TABLE IV

OBJECT DISTANCE EVALUATION ON NUSCENES. THE ALGORITHMS EVALUATED ARE STATE-OF-THE-ART MONOCULAR DEPTH ESTIMATION METHODS: MONODEPTH2 (MD2) AND BTS. DEPTH ERRORS WERE COMPUTED FOR THE OBJECT CLASSES WITH ENOUGH INSTANCE IN THE TEST SPLIT. BOTH SRE AND RMSE ARE EXPRESSED IN METERS.

	RE		SRE		RMSE		logRMSE		a_1		a_2		a_3	
Object class	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS	MD2	BTS
Car	0.346	**0.218**	6.853	**2.144**	10.42	**6.862**	0.448	**0.278**	0.546	**0.708**	0.736	**0.88**	0.92	**0.949**
Person	0.501	**0.384**	8.312	**3.91**	9.291	**7.858**	0.531	**0.449**	0.438	**0.492**	0.679	**0.717**	0.803	**0.839**
Bus	0.448	**0.228**	11.837	**2.226**	13.929	**7.811**	0.448	**0.274**	0.465	**0.644**	0.729	**0.891**	0.848	**0.958**
Truck	0.324	**0.218**	6.803	**2.091**	11.425	**7.263**	0.378	**0.26**	0.574	**0.674**	0.793	**0.902**	0.887	**0.964**
Motorcycle	0.284	**0.245**	1.671	**1.43**	4.509	**3.917**	0.32	**0.288**	0.512	**0.658**	0.868	**0.869**	0.935	**0.946**

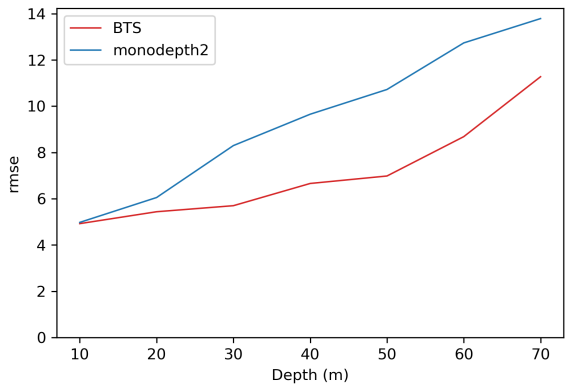

Fig. 2. Our quantitative RMSE results for the car object class over distance ranges on the KITTI dataset. RMSE is expressed in meters.

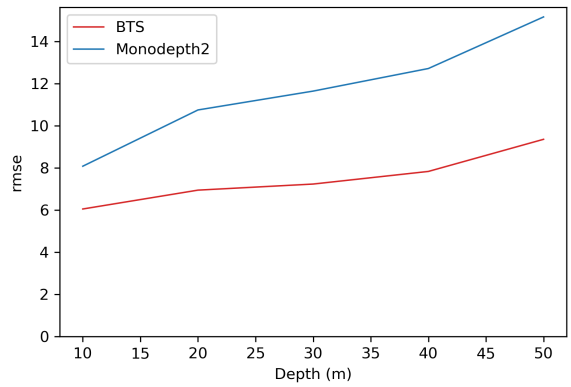

Fig. 3. Our quantitative RMSE results for the car object class over distance ranges on the NuScenes dataset. RMSE is expressed in meters.

makes learning the depth harder for the CNNs whereas the surrounding environment is less variable making it easier for the methods to learn the depth. We can see that errors can be as high as 2 time the global error for persons and cars, and it must be in mind if these methods are used to predict the distance of an object. Finally we can see that our results over the NuScenes dataset have higher errors than on KITTI, this can be explained by the differences in the training between the two datasets and NuScenes having more challenging scenes for depth estimation. For example, some scenes have been acquired in a raining weather which have reflections due to the wet road. Scenes have also been captured during nighttime with poor visibility. These scenes have a much higher error than those with good visibility and it contributes to increase the mean errors used for calculating the global error. By combining our two evaluation protocols, we also computed the evolution of the error for objects like cars over distance ranges in Figures 2 and 3. Such comparative results can be used to evaluate the fitness of a depth estimation method to a particular scenario in road environments. For example for driving in regular road where we assume that the vehicle is travelling at 90km/h the method have to be accurate up to 60m (safety distance between two vehicles at that speed).

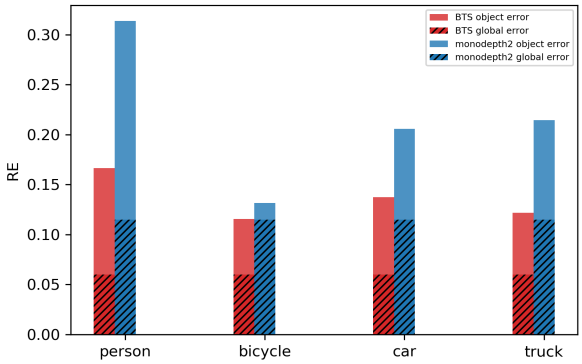

Fig. 4. Our RE results of BTS and Monodepth2 for different object classes compared to the global RE (hashed) on KITTI.

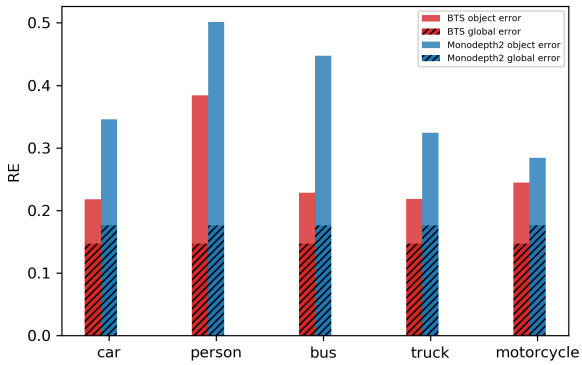

Fig. 5. Our RE results of BTS and Monodepth2 for different object classes compared to the global RE (hashed) on NuScenes.

V. Conclusions

We presented in this paper a novel depth evaluation protocol better suited for autonomous driving applications in road scenes as well as an evaluation of BTS and Monodepth2, two state-of-the-art monocular depth estimation methods, using our new protocols. For depth evaluation protocols, we proposed a protocol over distance ranges allowing to evaluate the evolution of the depth accuracy over the distance and a protocol for evaluating the depth predictions of objects by using the object masks generated using Mask-RCNN, one of the best object detector. We then performed an evaluation of an unsupervised method and an supervised one with Monodepth2 and BTS on two large scale datasets for road scenes, KITTI and NuScenes. The comparatives results of

the evaluation shows that BTS have better performance on all aspects than Monodepth2. We also showed that object depth estimation errors were significantly higher than the errors on the whole image for both methods. However while we offer a comprehensive evaluation of depth estimation over road environments, more methods including stereo-based algorithms could also be evaluated for a complete comparative study. We also aim to evaluate depth estimation algorithms on railway environments but due to the lack of a public dataset with camera images and depth ground truths from a LiDAR, we will need to acquire our own dataset for this task. Thus, we propose to develop an acquisition system including a stereoscopic camera and a LiDAR so that we can collect our own dataset in the railway environment.

ACKNOWLEDGMENT

This research is supported by SEGULA Technologies and M2SINUM project (This project is co-financed by the European Union with the European regional development fund (ERDF, 18P03390/18E01750/18P02733) and by the Haute-Normandie Regional Council via the M2SINUM project). We would like to thank SEGULA Technologies for their collaboration and the engineers of Autonomous Navigation Laboratory of IRSEEM for their support. This work was performed in part on computing resources provided by CRIANN (Centre Régional Informatique et d'Applications Numériques de Normandie, Normandy, France).

REFERENCES

[1] H. Mukojima, D. Deguchi, Y. Kawanishi, I. Ide, H. Murase, M. Ukai, N. Nagamine, and R. Nakasone, "Moving camera background-subtraction for obstacle detection on railway tracks," in *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 3967–3971, IEEE, 2016.

[2] S. Yanan, Z. Hui, L. Li, and Z. Hang, "Rail surface defect detection method based on yolov3 deep learning networks," in *2018 Chinese Automation Congress (CAC)*, pp. 1563–1568, IEEE, 2018.

[3] J. Palacín, T. Pallejà, M. Tresanchez, R. Sanz, J. Llorens, M. Ribes-Dasi, J. Masip, J. Arno, A. Escola, and J. R. Rosell, "Real-time tree-foliage surface estimation using a ground laser scanner," *IEEE transactions on instrumentation and measurement*, vol. 56, no. 4, pp. 1377–1383, 2007.

[4] B. Kang, S.-J. Kim, S. Lee, K. Lee, J. D. Kim, and C.-Y. Kim, "Harmonic distortion free distance estimation in tof camera," in *Three-Dimensional Imaging, Interaction, and Measurement*, vol. 7864, p. 786403, International Society for Optics and Photonics, 2011.

[5] R. Khemmar, M. Gouveia, B. Decoux, and J.-Y. Ertaud, "Real time pedestrian and object detection and tracking-based deep learning. application to drone visual tracking," in *International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision*, 01 2019.

[6] Z. Chen, R. Khemmar, B. Decoux, A. Atahouet, and J.-Y. Ertaud, "Real time object detection, tracking, and distance and motion estimation based on deep learning: Application to smart mobility," in *2019 Eighth International Conference on Emerging Security Technologies (EST)*, pp. 1–6, IEEE, 2019.

[7] A. Mauri, R. Khemmar, B. Decoux, N. Ragot, R. Rossi, R. Trabelsi, R. Boutteau, J.-Y. Ertaud, and X. Savatier, "Deep learning for real-time 3d multi-object detection, localisation, and tracking: Application to smart mobility," *Sensors*, vol. 20, no. 2, p. 532, 2020.

[8] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[9] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," *arXiv preprint arXiv:1903.11027*, 2019.

[10] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3431–3440, June 2015.

[11] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, p. 2481–2495, Dec 2017.

[12] J. H. Lee, M.-K. Han, D. W. Ko, and I. H. Suh, "From big to small: Multi-scale local planar guidance for monocular depth estimation," *arXiv preprint arXiv:1907.10326*, 2019.

[13] I. Alhashim and P. Wonka, "High quality monocular depth estimation via transfer learning," *arXiv preprint arXiv:1812.11941*, 2018.

[14] G. Huang, Z. Liu, L. v. d. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[15] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012.

[16] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 270–279, 2017.

[17] Y. Luo, J. Ren, M. Lin, J. Pang, W. Sun, H. Li, and L. Lin, "Single view stereo matching," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018.

[18] F. Tosi, F. Aleotti, M. Poggi, and S. Mattoccia, "Learning monocular depth estimation infusing traditional stereo knowledge," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9799–9809, 2019.

[19] C. Godard, O. Mac Aodha, M. Firman, and G. Brostow, "Digging into self-supervised monocular depth estimation," *arXiv preprint arXiv:1806.01260*, 2018.

[20] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1851–1858, 2017.

[21] V. Casser, S. Pirk, R. Mahjourian, and A. Angelova, "Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 8001–8008, 2019.

[22] H. Zhan, R. Garg, C. S. Weerasekera, K. Li, H. Agarwal, and I. M. Reid, "Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018.

[23] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," 2017.

[24] T. Koch, L. Liebel, F. Fraundorfer, and M. Körner, "Evaluation of cnn-based single-image depth estimation methods," *CoRR*, vol. abs/1805.01328, 2018.

[25] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE international conference on computer vision*, pp. 2650–2658, 2015.