A Comparative Study of Deep Learning-based Depth Estimation Approaches: Application to Smart Mobility

Antoine Mauri, Redouane Khemmar, Benoit Decoux, Tahar Ben Moumen, Madjid Haddad, Rémi Boutteau

A Comparative Study of Deep Learning-based Depth Estimation Approaches: Application to Smart Mobility

Antoine Mauri, Redouane Khemmar, Benoit Decoux, Tahar Benmoument, Madjid Haddad, Rémi Boutteau

2021 8th International Conference on Smart Computing and Communications (ICSCC)

A Comparative Study of Deep Learning-based Depth Estimation Approaches: Application to Smart Mobility

1st Antoine Mauri
Normandie Univ, UNIROUEN
ESIGELEC, IRSEEM
76000 Rouen, France
antoine.mauri@esigelec.fr

2nd Redouane Khemmar
Normandie Univ, UNIROUEN
ESIGELEC, IRSEEM
76000 Rouen, France
redouane.khemmar@esigelec.fr

3rd Benoit Decoux
Normandie Univ, UNIROUEN
ESIGELEC, IRSEEM
76000 Rouen, France
benoit.decoux@esigelec.fr

4th Tahar Benmoumen
Normandie Univ, UNIROUEN
ESIGELEC, IRSEEM
76000 Rouen, France
benmoumentahar0@gmail.com

5th Madjid Haddad
SEGULA Technologies
92000 Nanterre, France
Madjid.HADDAD@segula.fr

6th Rémi Boutteau
Normandie Univ, UNIROUEN
UNILEHAVRE, INSA Rouen, LITIS
76000 Rouen, France
remi.boutteau@univ-rouen.fr

***Abstract*—In autonomous vehicle systems, the quality of scene perception is of great importance for security preoccupation in road environments. In this context, an accurate localization of potential obstacles is one of the most challenging tasks. In recent years, substantial progress has been made in the field of depth estimation for detection purposes with the spread of methods relying on deep learning with monocular or stereoscopic camera(s). These two families of approaches did show an upstanding yet inconsistent performance in different road scenes circumstances. A deep understanding and comparison of these approaches is required to allow the community an easier assessment, which breeds to more adequate choice for their own systems. In this paper, we propose a comparative study of state-of-the-art deep learning depth estimation methods using monocular and stereoscopic cameras. The evaluation is performed on road environment over the challenging KITTI dataset.**

***Index Terms*—Depth estimation, deep learning evaluation, computer vision, smart mobility, monocular and stereoscopic approaches, KITTI dataset.**

I. Introduction

Accurate and reliable depth estimation is essential for the perception of the environment in front of the vehicle and can drastically increase safety by estimating the distance between the vehicle and a potential obstacles. As demonstrated in [1] and [2], technology also plays an important role in improving the competitiveness of road transport. Nevertheless, many challenges remain to be addressed before this technology becomes fully operational. The measurement of the distance to objects is mainly based on different sensors: Radar, Lidar [3] or time-of-flight camera [4]. However, most of these sensors are very expensive and can be cumbersome. In this work, we focus instead on the use of cameras for depth estimation. The most common method for this task is to use a stereoscopic sensor composed of a pair of calibrated cameras. Recently, methods based on Convolutional Neural Networks (CNNs) have been explored in depth estimation and showed good performance for stereoscopic images as well as single images. The main advantage of these methods is the reduction in the cost of materials. Lidar and Radar are replaced by a standard camera (stereo or mono) which is easy to integrate and low-cost. Many approaches have been proposed in the literature, however, there is still a lack of studies that have been done to evaluate these methods under realistic environments like road or rail traffic environment.

The main contribution of our paper is the proposition of a comparative study of deep learning approaches for depth estimation from either a single or stereoscopic images. The objective of this study is to offer qualitative and quantitative evaluation of the state-of-the-art methods for road environments. This work is in line with the work we have carried out on the perception of the environment for the autonomous vehicle [5], [6]. The remainder of this paper is organized as follows. Section I introduces the paper. In section II, we review the depth estimation approaches which are evaluated in this paper and the already existing comparative studies for depth estimation. Section III presents in more detail the methodology used in our evaluation. The experimental results of our evaluation are presented in section IV. Finally, we draw our conclusions and future directions in section V.

II. Depth Estimation Methods

A. Overview

In computer vision applications, depth estimation is a key task which is designed to estimate depth of objects captured from 2D images. Depth estimation task requires as input data only 2D RGB image and generates as output data a depth image made of the distances between objects detected in the scene and the viewpoint of the camera. An example of depth images can be found in Figure 1.

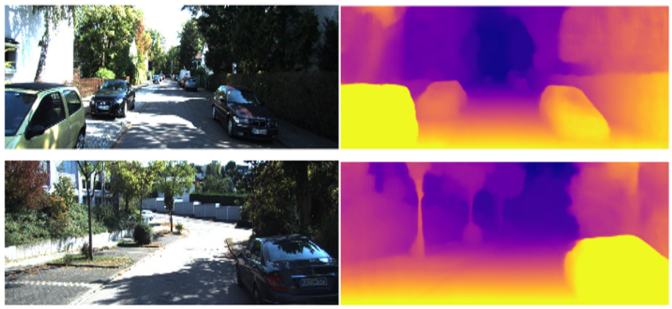

Fig. 1. Depth images from KITTI Dataset [7]. RGB images can be found on the left and depth images on the right.

One of the main applications of depth estimation task in computer vision is the autonomous vehicle, and more broadly smart mobility. With regards to such application, we need to enhance the quality of the environment perception of road mobility by improving not only objects detection, but also their localization and tracking. This is why depth estimation task represents a crucial step after our data acquisition and processing platform for smart mobility. In this section, we will present different methods of depth estimation based on deep learning which we consider in our comparative study.

B. Monocular Depth Estimation Methods

In [8], a fully convolutional architecture-based CNN for depth estimation from RGB images of a given scene is presented. Modeling of equivocal correspondence between monocular images and depth maps is performed using residual learning. The optimization of the model is carried out through the RHL (Reverse Huber Loss). The approach runs in real-time on both images and videos. In [9], Zhou et al. present an unsupervised learning framework dedicated to both single-view CNN monocular depth estimation task and estimation of the camera ego-motion using unstructured video sequences. Both single-view depth and multi-view pose networks are used in their framework. Based on Dispnet [10], this method leverages an encoder-decoder architecture with skip connections and multi-scale side predictions. The model is validated on the Cityscape dataset [11].

In Monodepth2 [12], authors propose a trained CNN for single frame depth estimation without supervision from a ground-truth. The end-to-end unsupervised monocular training is performed using a training loss enforcing the depth consistency from the left to the right. The CNN architecture is also inspired by DispNet. Higher resolution details are recovered by using the skip connections between the encoder's activation blocks. Two disparity maps are predicted : left-to-right and right-to-left. The model is validated on the KITTI dataset [7]. In [13], Casser et al. present the unsupervised learning for depth scenes and ego-motion, with only the supervision from monocular videos. The learning process uses geometric structure in order to involve modeling the scene and the individual objects. With an RGB image sequences as input and pre-computed segmentation masks, the models predicts the transformation vectors per object in the 3D environment. This model is validated on the KITTI dataset.

The approach proposed in [14] proposes an unsupervised method for ego-motion and depth estimation from a single frame input. The motions of objects are predicted in 3D. The model's outputs are the individual warping from moving objects and the camera's ego-motion. Results of the approach are also evaluated and validated over the KITTI dataset.

In [15], a Neural Network is presented for the reconstruction of a piecewise planar depth map from a single frame. A set of plane parameters and segmentation masks are produced by the method, called PlaneNet, from a single frame. The network can use a loss defined by the probalistic segmentation to predict a depth map for non-planar surfaces. PlaneNet has three prediction tasks: plane parameters, non-planar depth maps, and segmentation masks. The PlaneNet model is evaluated over the NYUv2 dataset [16] through depth accuracy comparison between different approaches like Eigen-VGG [17], SURGE [18], and FCRN [8]. In BTS (Big-To-Small) approach [19], the full resolution depth is obtained by merging the outputs of different intermediate layers of the decoder. The proposed CNN architecture allows to bypass the problem caused by the bottleneck of the encoder-decoder architecture to obtain a full resolution depth map. This architecture is currently among the best performing methods on the KITTI dataset depth estimation benchmark.

C. Stereoscopic Methods

1) Pyramid Stereo Matching Network: A novel network is introduced in [20] called Pyramid Stereo Matching Network (PSMNet) which exploits global context information in stereo matching. In order to learn and expand the receptive fields, Spatial Pyramid Pooling (SPP) [21] [22] and dilated convolution [23] [24] have been used. Expressly, PSMNet expands pixel-level features to region-level features with distinct scales of receptive fields; the ensuing combined global and local features are used to shape the cost volume for dependable disparity estimation. The authors also tend to design a stacked hourglass network 3D CNN-based along an intermediate supervision for cost volume supervision. The stacked hourglass-shaped 3D CNN processes cost volume in a top-down and bottom-up manner to enhance the use of global contextual information. The PSMNet includes two main modules: SPP and 3D CNN.

SPP: It is troublesome to see the context relationship alone from pixel intensities. As a result, image features made with

object context information will enhance the matching process especially for ill-posed regions. Therefore, PSMNet incorporates the SPP module to learn the connection between associate object and its sub-regions, which helps the integration of graded context information. In [21], SPP was mainly proposed to avoid the fixed-size constraint of CNNs. The different levels of the features map generated by the SPP are used for the classification being passed through a fully connected layer.

3D CNN: By leveraging the different levels of features, the SPP modules, manages to improve the stereo matching in the network. For the combination of information from the the features along the disparity and spatial dimensions, authors propose two 3D CNN structures for cost volume regularization: the basic and the stacked hourglass architectures. For the first one, the network is simply built using residual blocks. In stacked hourglass architecture, the network consists on repeating top-down/bottom-up processing in conjunction with intermediate supervision.

2) Group-wise Correlation Stereo Network: In [25], Group-wise correlation Network (GwcNet) was introduced in order to estimate disparity maps using stereo data. The approach include group-wise correlation for building up the cost volumes.For improving the performances and reducing the number of parameters of the network, group correlation volumes are used to deliver suitable corresponding features for the 3D aggregation network. Reported experiments show that when the computational cost is bounded, their model achieves larger gain than similar state-of-the-art networks. They also improved the stacked hourglass networks by improving the performance and reducing the inference time. Therefore, as shown earlier, the group-wise correlation goal is to tackle the drawbacks such as losing information in the full correlation or the huge requirement of parameters setting in the concatenation volume.

Multi-level features are extracted and concatenated to construct high-dimensional feature representations f_l, f_r for a pair of images from a stereoscopic camera. The features are then divided into groups based on channel dimension, and the corresponding left and right feature groups are correlated with each other across all levels of disparity to obtain correlation maps by group. Finally, all correlation maps are transformed into a 4D cost volume. Features can be processed as structured vector groups [26], so correlation maps for a specific group can be considered as a proposed matching cost. Therefore, the power of traditional cross-correlation matching cost is increased and the provided similarity measures is enhanced. GwcNet [25] enhances PSMNet [20] with group-wise correlation cost volume and improvement for the stacked hourglass networks. For PSMNet, the mapping costs for the concatenated features must be learned from scratch by the 3D aggregation network, resulting in increased parameters and computational costs. On the other hand, using full correlation [10] is an efficient mean to measure similarities between features, but at the risk of losing informations. The stacked hourglass architecture proposed in PSMNet allows to better learn the context features, yet, in order to improve the inference speed, authors of the GWCNet have made two main changes on the structure of the 3D aggregation. Firstly, they added one more auxiliary output module for the features of the pre-hourglass module which improves the network's learning of features at lower layers for an improved final prediction. Secondly, residual connections between different modules are removed, so that the outputs of auxiliary modules can be removed during inference to reduce the computational cost.

Taking into consideration all the cited pros and cons of both GwcNet and PSMNet, we choose to evaluate these two networks and compare their performance on the KITTI dataset. The corresponding algorithms and pre-trained models are publicly available so that the interested readers can try easily these networks on their own data.

D. Evaluation of Depth Estimation Methods

While some work has been done in terms of comparative study of depth estimation methods for either stereo or monocular camera, few work features a comparative study of both monocular and stereo. [27] and [28] present a comprehensive survey of stereo-based depth estimation as well as an in-depth evaluation of the methods. [29] offers a new set of evaluation protocols devoted to single image depth estimation in order to better assess the performance of the proposed methods. It provides mainly an evaluation of multiple monocular methods in indoor environments. Also, in [30] a comparison and evaluation of multiple encoder architectures for depth estimation is proposed. None of these papers went for comprehensive comparison and evaluation of both stereoscopic and monocular depth estimation in road environment, which is the aim of this paper.

III. COMPARATIVE STUDY DEPTH ESTIMATION APPROACHES BASED ON DEEP LEARNING

A. Metrics used for Depth Evaluation

To be aligned with the state of the art, we choose to use the common metrics used to evaluate depth estimation approaches: Relative Error (RelErr), Squared Relative Error (SqRel), Root Mean Squared Error (RMSE), and Logarithmic Root Means Squared Error (logRMSE). These metrics give an overall assessment of a method's performance in the entire tested image. We present expressly each metric in what follows. We note by *p* the depth prediction, *gt* as its corresponding ground truth of size *N*.

Relative Error (RelErr) is detailed in Equation (1).

$$RelErr = \frac{1}{N}\sum_{u}\sum_{v}\frac{|gt_{u,v} - p_{u,v}|}{gt_{u,v}} \tag{1}$$

Squared Relative Error (SqRelErr) is detailed in Equation (2).

$$SqRelErr = \frac{1}{N}\sum_{u}\sum_{v}\frac{|gt_{u,v} - p_{u,v}|^2}{gt_{u,v}} \tag{2}$$

Root Mean Squared Error (RMSE) equation can be found in Equation (3).

$$RMSE = \sqrt{\frac{1}{N}\sum_{u}\sum_{v}(gt_{u,v} - p_{u,v})^2} \tag{3}$$

Logarithmic Root Mean Squared Error (logRMSE) is detailed in Equation (4).

$$logRMSE = \sqrt{\frac{1}{N}\sum_{u}\sum_{v}(log(gt_{u,v}) - log(p_{u,v}))^2} \quad (4)$$

B. KITTI Dataset

Studied methods for both single images or stereo images, have been pretrained and evaluated on the KITTI dataset. KITTI is one of the most widely used dataset for environment perception in road environment. This dataset offers synchronized images from a stereoscopic camera with a multitude of sensors such as a velodyne 3D laser scanner and a high precision GPS/IMU navigation system. The acquisitions were made in high density real-world traffic situations and thus is regarded as a challenging dataset for environment perception methods such as depth estimation. For ground truth, we use synchronized and calibrated data from both the velodyne and camera sensors.

IV. EXPERIMENTAL RESULTS

A. Evaluation of Single Image-based Methods

We used the pretrained models published by the authors of BTS and Monodepth2 networks for the evaluation on the KITTI dataset. BTS was trained with a resolution of 704×352 on Eigen's training split [17] with dense ground truth. Monodepth2 was trained using its unsupervised monocular training on Zhou's training split [9] with an input resolution of 1024×320. Both methods have been evaluated on the Eigen test split. The results can be found in Table I.

TABLE I
MONOCULAR DEPTH EVALUATION ON KITTI. EVALUATED METHODS INCLUDE MONODEPTH2 (MD2) AND BTS, TWO STATE-OF-THE-ART MONOCULAR DEPTH ESTIMATION METHOD. THE SQREL AND RMSE ARE EXPRESSED IN METERS.

	RelErr	SqRel	RMSE	logRMSE
Monodepth2	0.115	0.882	4.701	0.190
BTS	**0.060**	**0.249**	**2.798**	**0.096**

The results show that BTS is in overall performing better on all error metrics than Monodepth2. One of the reasons why BTS performs better than Monodepth2 is probably because it was trained with supervisory depth information while Monodepth2 was trained in an unsupervised way. Monodepth2 used sequences of images for the unsupervised training and used a model to learn the ego-motion of the camera for supervision. This results in a lot of approximations during training. Moreover, BTS model is much deeper and computational heavy than Monodepth2.

B. Evaluation of Stereoscopic Image-based Methods

The results of the stereoscopic methods on the KITTI dataset shows that GWcNet is significantly better than PSMNet. Qualitative results can be found in Figure 2 and quantitative results are show in Table II.

Through experimental results of the two compared stereo approaches, we can explain the difference in terms of performance between the GwcNet and the PSMNet by the fact that

TABLE II
STEREOSCOPIC DEPTH EVALUATION ON KITTI. EVALUATED METHODS INCLUDE GWCNET AND PSMNET, TWO STATE-OF-THE-ART STEREOSCOPIC DEPTH ESTIMATION METHOD. THE SQREL AND RMSE ARE EXPRESSED IN METERS.

	RelErr	SqRel	RMSE	logRMSE
GWCNET	**0.018**	**0.048**	**0.981**	**0.042**
PSMNET	0.032	0.061	1.139	0.056

the feature aggregation functionalities are making the network more robust against challenging road scenes (like textureless objects and regions, sudden illumination changes, etc.) which are more present in the GcwNet owing to the group-wise correlation highly recommended in stereo-like networks. We

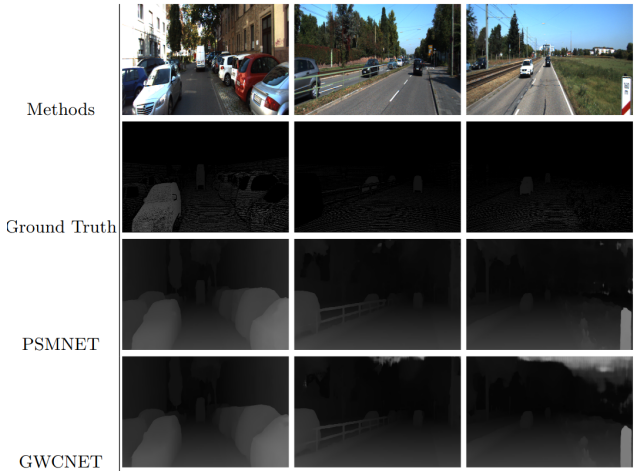

Fig. 2. Qualitative results of stereo-based depth estimation methods (GWCNet and PSMNet) on samples from KITTI.

are also currently working on testing the stereoscopic methods presented in this paper on our own developed stereo camera, this would allow us to use the methods on railway environment for the autonomous train where little work has been done compared to the autonomous car.

V. CONCLUSION AND FUTURE WORK

In this paper, we have presented a comprehensive comparative study of deep learning-based depth estimation methods for both monocular and stereoscopic camera. The challenges of this work is the reuse and the comparison of these state-of-the-art methods on complex road environments through the evaluation over the KITTI dataset. The real-world traffic condition from this dataset gives a good assessment of how a method can perform in realistic conditions. Our experimental evaluation showed that BTS and GWCNet offer the best performances for monocular and stereoscopic cameras, respectively. Our work also shows that stereoscopic methods have a greater accuracy than monocular-based methods. Thus, we offer a comprehensive evaluation of depth estimation over road environments and we aim at trying these approaches on rail environments. No evaluation can be done on this environment due to the lack of a public dataset that includes

depth ground truth. Acquiring our own railway dataset with an acquisition system including a stereoscopic and a LiDAR sensors (through our IRSEEM autonomous platform made of a Renault Espace car with an instrumented roof-box containing multisensors fusion system: three perspectives cameras, one stereoscopic camera, one HDL-64/VLP16 LIDAR, one RTK GPS and one LANDINS Central Inertial Unit) should push forward the state of the art in this field.

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 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] 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.

[6] A. Mauri, R. Khemmar, R. Boutteau, B. Decoux, J. Y. Ertaud, and M. Haddad, "A new evaluation approach for deep learning-based monocular depth estimation methods," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1–6, Sep. 2020.

[7] 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.

[8] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *2016 Fourth international conference on 3D vision (3DV)*, pp. 239–248, IEEE, 2016.

[9] 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.

[10] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4040–4048, 2016.

[11] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

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

[13] 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.

[14] V. Casser, S. Pirk, R. Mahjourian, and A. Angelova, "Unsupervised monocular depth and ego-motion learning with structure and semantics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.

[15] C. Liu, J. Yang, D. Ceylan, E. Yumer, and Y. Furukawa, "Planenet: Piece-wise planar reconstruction from a single rgb image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2579–2588, 2018.

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

[17] 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.

[18] P. Wang, X. Shen, B. Russell, S. Cohen, B. Price, and A. L. Yuille, "Surge: Surface regularized geometry estimation from a single image," in *Advances in Neural Information Processing Systems*, pp. 172–180, 2016.

[19] 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.

[20] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5410–5418, 2018.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *European Conference on Computer Vision*, pp. 346–361, 2014.

[22] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2881–2890, 2017.

[23] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *arXiv preprint arXiv:1606.00915*, 2016.

[24] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[25] X. Guo, K. Yang, W. Yang, X. Wang, and H. Li, "Group-wise correlation stereo network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3273–3282, 2019.

[26] Y. Wu and K. He, "Group normalization," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

[27] H. Laga, "A survey on deep learning architectures for image-based depth reconstruction," *arXiv preprint arXiv:1906.06113*, 2019.

[28] H. Laga, L. V. Jospin, F. Boussaid, and M. Bennamoun, "A survey on deep learning techniques for stereo-based depth estimation," *arXiv preprint arXiv:2006.02535*, 2020.

[29] T. Koch, L. Liebel, F. Fraundorfer, and M. Korner, "Evaluation of cnn-based single-image depth estimation methods," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 0–0, 2018.

[30] M. Aladem, S. Chennupati, Z. El-Shair, and S. A. Rawashdeh, "A comparative study of different cnn encoders for monocular depth prediction," in *2019 IEEE National Aerospace and Electronics Conference (NAECON)*, pp. 328–331, IEEE, 2019.