This full text paper was peer-reviewed at the direction of IEEE Instrumentation and Measurement Society prior to the acceptance and publication.

PT-ResNet: Perspective Transformation-Based Residual Network for Semantic Road Image Segmentation

Rui Fan[1]*, Yuan Wang[1]*, Lei Qiao[2], Ruiwen Yao[2], Peng Han[2], Weidong Zhang[2], Ioannis Pitas[3], Ming Liu[1]
[1]Robotics Institute, the Hong Kong University of Science and Technology, Hong Kong.
[2]Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China.
[3]Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece.
eeruifan@ust.hk, ywangeq@connect.ust.hk, qiaolei2008114106@gmail.com, yaoruiwen88@foxmail.com,
han_ipac@sjtu.edu.cn, wdzhang@sjtu.edu.cn, eelium@ust.hk, pitas@csd.auth.gr

***Abstract*—Semantic road region segmentation is a high-level task, which paves the way towards road scene understanding. This paper presents a residual network trained for semantic road segmentation. Firstly, we represent the projections of road disparities in the v-disparity map as a linear model, which can be estimated by optimizing the v-disparity map using dynamic programming. This linear model is then utilized to reduce the redundant information in the left and right road images. The right image is also transformed into the left perspective view, which greatly enhances the road surface similarity between the two images. Finally, the processed stereo images and their disparity maps are concatenated to create a set of 3D images, which are then utilized to train our neural network. The experimental results illustrate that our network achieves a maximum F1-measure of approximately 91.19%, when analyzing the images from the KITTI road dataset.**

I. Introduction

Autonomous driving technology has been developing rapidly, since Google launched its self-driving car project in 2009 [1]. In recent years, industry titans, such as Waymo and Tesla, race to commercialize autonomous vehicles (AVs) [2], [3]. However, a number of high-profile experimental accidents that occurred in the last year and have called into question whether the autonomous driving technology is mature enough for employment [4]. Therefore, most researchers believe that in the next few years the research on autonomous driving should focus on developing advanced driver assistance systems (ADASs) [5], such as lane marking detection [6], road surface 3D reconstruction [7], 2D/3D object detection [8], localization and mapping [9], [10], etc.

Visual environment perception (VEP) is a key component of ADAS [11], [12]. After learning from a large amount of labeled training data, VEP can extract useful road environment information, e.g., free space areas and pedestrians, from road images [13], semantic image region segmentation can provide useful information by partitioning an image into semantically meaningful regions and classifying them into one of the predefined categories [14]. State-of-the-art semantic segmentation algorithms are generally based on fully convolutional networks (FCNs) [14], which are an extension of convolutional neural network (CNN). FCNs utilize classical CNNs to learn image feature representations, but the input images can be of any sizes. FCNs perform image upsampling to produce a probability mask with the same size as the input image [15].

FCN-LC [15] is a classical FCN used for semantic road image segmentation. FCN-LC utilizes a network-in-network architecture [15] to learn road region segmentation from labeled training image data. This allows fast inference, even for large contextual image window sizes [15]. In addition, in recent years, a number of conditional random field (CRF)-based neural networks, e.g., PGM-ARS [16], Hybrid [17] and StixelNet [18] have been proposed for semantic road image segmentation. PGM-ARS [16] and StixelNet [18] were trained using monocular images, while Hybrid [17] also employed the 3D road scene information acquired using LiDAR for training. Furthermore, stereo vision [19], [20] was used to improve road segmentation performance. For example, a so-called BM neural network [19] selects a region of interest (ROI) in an image, by analyzing the v-disparity information. Such ROI information greatly minimizes the number of incorrectly segmented pixels. Furthermore, a so-called HistonBoost network [20] post-processes such ROIs using watershed transformation and morphological filtering [21]. In this paper, we draw on the success of [19], [20] and present a perspective transformation (PT)-based deep convolutional network for road semantic segmentation. It is designed using the residual network (ResNet) from DeepLab [22]. The structure of our proposed network is shown in Fig. 1.

The rest of the paper is organized as follows: Section II introduces the proposed PT-ResNet. In Section III, the experimental results are illustrated and the performance of the proposed approach is evaluated. Section IV contains conclusion and some recommendations for future work.

II. Methodology

A. Training Data Pre-Processing and Generation

3D information can greatly enhance VEP robustness [23]. In this paper, the proposed semantic segmentation method focuses entirely on the road surface, which can be treated as a ground plane. According to the perspective transformation algorithm presented in [24], a right image can be transformed into its left view using the disparity projection model. This can greatly enhance the similarity of the road surface between the stereo images [24]. Therefore, in this paper, we first utilize PSMNet [25] to estimate dense disparity maps (see Fig. 2). A

*This two authors are joint first authors.

978-1-7281-3868-3/19/$31.00 ©2019 IEEE

Authorized licensed use limited to: Aristotle University of Thessaloniki. Downloaded on September 29,2023 at 11:05:14 UTC from IEEE Xplore. Restrictions apply.

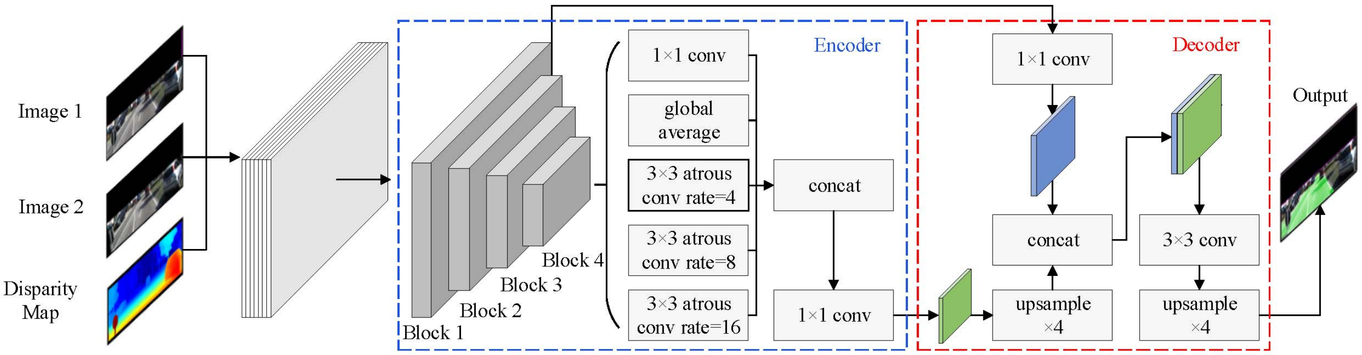


Fig. 1. PT-ResNet structure.

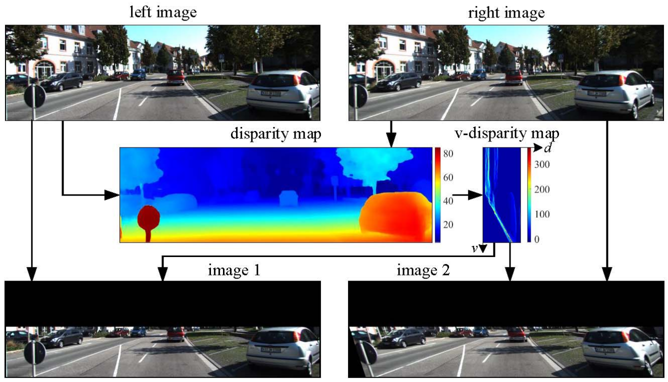


Fig. 2. Training data pre-processing and generation.

v-disparity map is then created by computing the histograms $\hat{p}(d, v)$ of each horizontal row v of the disparity map [26]. To find the path corresponding to the road disparity projection in the v-disparity map, we utilize dynamic programming (DP) to search for every possible solution [27]:

$$E(d, v) = -\hat{p}(d, v) + \min_{\tau=0}^{\tau_{\text{max}}}[E(d + 1, v - \tau) - \lambda\tau], \quad (1)$$

where $\hat{p}(d, v)$ represents the histogram value at (d, v) in the v-disparity map, λ is a smoothness term, τ_{max} is the maximum search range [28]. E represents the energy of each possible solution. The path corresponding to the road disparity projection is generally represented using a linear polynomial [24]:

$$f(v) = \alpha_0 + \alpha_1 v. \quad (2)$$

The vertical coordinate of the vanishing point, i.e., v_{py}, can be estimated using (2). As the vertical coordinates of the road pixels are always larger than v_{py}, the image region above the vanishing point can be removed from the left and right images (see Fig. 2). Then, we utilize our previous algorithm [24] to transform the perspective view of the right image. This algorithm improves the road surface similarity in the stereo images, but also distorts obstacles, such as vehicles and trees. Finally, the processed stereo images and the left disparity maps are concatenated together to generate a set of 3D images with seven channels, which are then utilized to train the neural network.

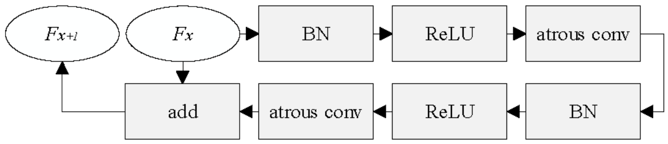


Fig. 3. The structure of each block unit in Fig. 1.

B. *PT-ResNet Structure*

In recent years, the encoder-decoder structure has been prevalently used in deep neural networks for semantic segmentation [29]. The encoder allows fast high-dimensional image feature map generation, while the decoder enables the network to recover sharp object boundaries [29]. In this paper, our network is designed following ResNet-101 used in DeepLab-v3+ [29]. The structure of our proposed network is shown in Fig. 1.

1) Encoder: In the encoder, the spatial dimension of the feature maps reduces gradually using four blocks, as shown in Fig. 1. The structure of each block is shown in Fig. 3, where BN denotes batch normalization, and ReLU denotes rectified linear unit. ReLU activation function is used to avoid overfitting during training. The parameter of ReLU is set to 0.5 in this paper. BN is a method used to normalize the input of each layer and overcome the internal covariate shift problem. As the stride in each block is set to 2, the output of the fourth block is 256 times smaller than the input of the first block. Furthermore, the baseline utilizes an atrous convolutional layer instead of a conventional convolutional layer. This allows us to enlarge the field-of-view of filters when interpolating the multi-scale context in the framework of spatial pyramid pooling and cascaded modules [30]. The output of each block can be computed by adding the output of the atrous convolutional layer to the input of the block, as shown in Fig. 3. The output of the fourth block feeds into five branches, as shown in Fig. 3. The baseline uses global average pooling to obtain the global image feature representations. In addition, three atrous convolutional layers with different rates are utilized to acquire multi-scale information. The rates depend entirely on the feature map produced by block 4, and they are set to 4, 8 and 16, respectively. Finally, the five branch output is concatenated and further compressed using a 1×1

Authorized licensed use limited to: Aristotle University of Thessaloniki. Downloaded on September 29,2023 at 11:05:14 UTC from IEEE Xplore. Restrictions apply.

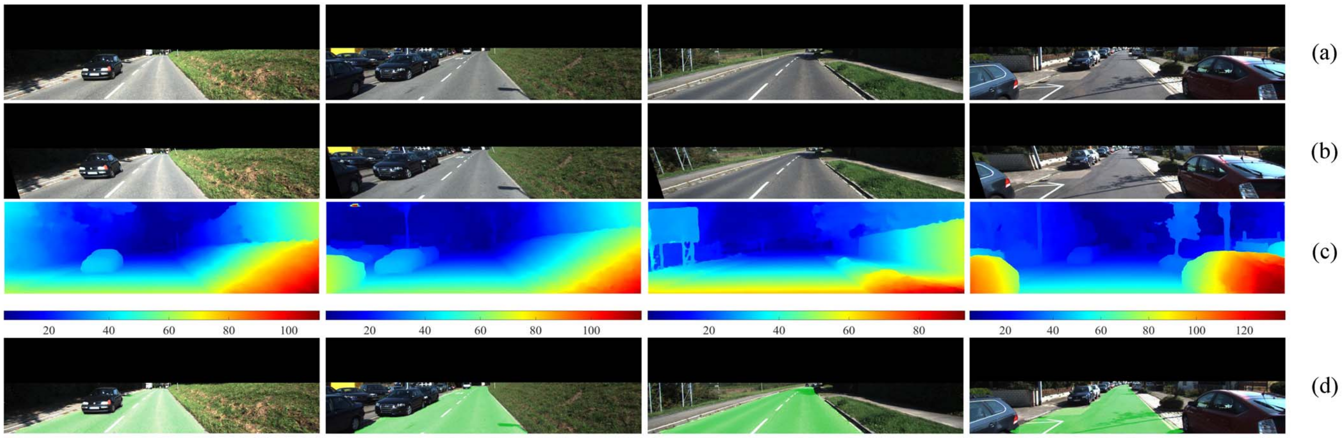


Fig. 4. Experimental results of road semantic segmentation (threshold is set to 0.9). The green areas in the fourth row are the segmented road surfaces. (a) Processed left images. (b) Transformed right images. (c) Disparity maps. (d) Segmentation results.

convolutional layer.

2) Decoder: In the decoder, the baseline applies skip connection to the feature map, which is produced by the second block. This greatly improves the details of local features in the high-level feature map. The low-level and high-level feature maps are then concatenated together. A probability map can be obtained after a feature map upsampling process. By finding the pixels whose probabilities are higher than our pre-set threshold, the semantic segmentation result can be obtained. Some examples of experimental results are shown in Fig. 4.

III. Experimental Results

In this section, we present our experimental results and evaluate the performance of the proposed method using the KITTI road dataset [31]. The dataset contains synchronized stereo road image pairs, 3D road scenery point clouds acquired using a Velodyne HDL-64E LiDAR, calibration parameters, and semantic region segmentation ground truth. The images in this dataset are grouped into three categorizes: urban unmarked (UU), urban marked (UM) and urban multiple marked (UMM). To quantify the accuracy of the proposed approach, a set of indicators, including maximum F1-measure (MaxF), average precision (AP), precision (PRE), recall (REC), false positive rate (FPR) and false negative rate (FNR), are computed and are publicly available on the KITTI road benchmark[1]. PT-ResNet training was conducted on an NVIDIA GTX 1080 Ti GPU (CUDA 9 and cnDNN v7). In the experiments, the learning rate, training step and batch size are set to 0.001, 30000 and 8, respectively. The approach was programmed in Python language. The runtime of segmenting an image from the KITTI dataset is around 3 seconds. In this section, we compare our method with FCN-LC [15], PGM-ARS [16], Hybrid [17], StixelNet [18], BM [19] and HistonBoost [20]. The comparisons of MaxF, AP, PRE, REC, FPR and FNR among these methods are shown in Fig. 5, where urban reflects the overall performance of UM, UMM and UU. It can be observed in Fig. 5(a) that our PT-ResNet method outperforms the others in terms of MaxF, it achieves a MaxF of approximately 91.91%, which is slightly higher than that achieved using FCN-LC (90.79%). Fig. 5(b) indicates that PT-ResNet performs better than other networks in terms of AP, as it achieves an AP of approximately 91.21%. However, FCN-LC performs slightly better than our network in terms of PRE and FPR (see Fig. 5(c) and 5(e)). The overall PRE and FPR we achieved is 90.78% and 5.13%, respectively. Additionally, PT-ResNet achieves an intermediate performance in terms of REC and FNR (see Fig. 5(d) and 5(f)), as the REC and FNR values obtained using our method is 91.60% and 8.40%, respectively. In general, the proposed PT-ResNet achieves the best overall performance and its ranking is higher than that of other CNNs.

IV. Conclusion and Future Work

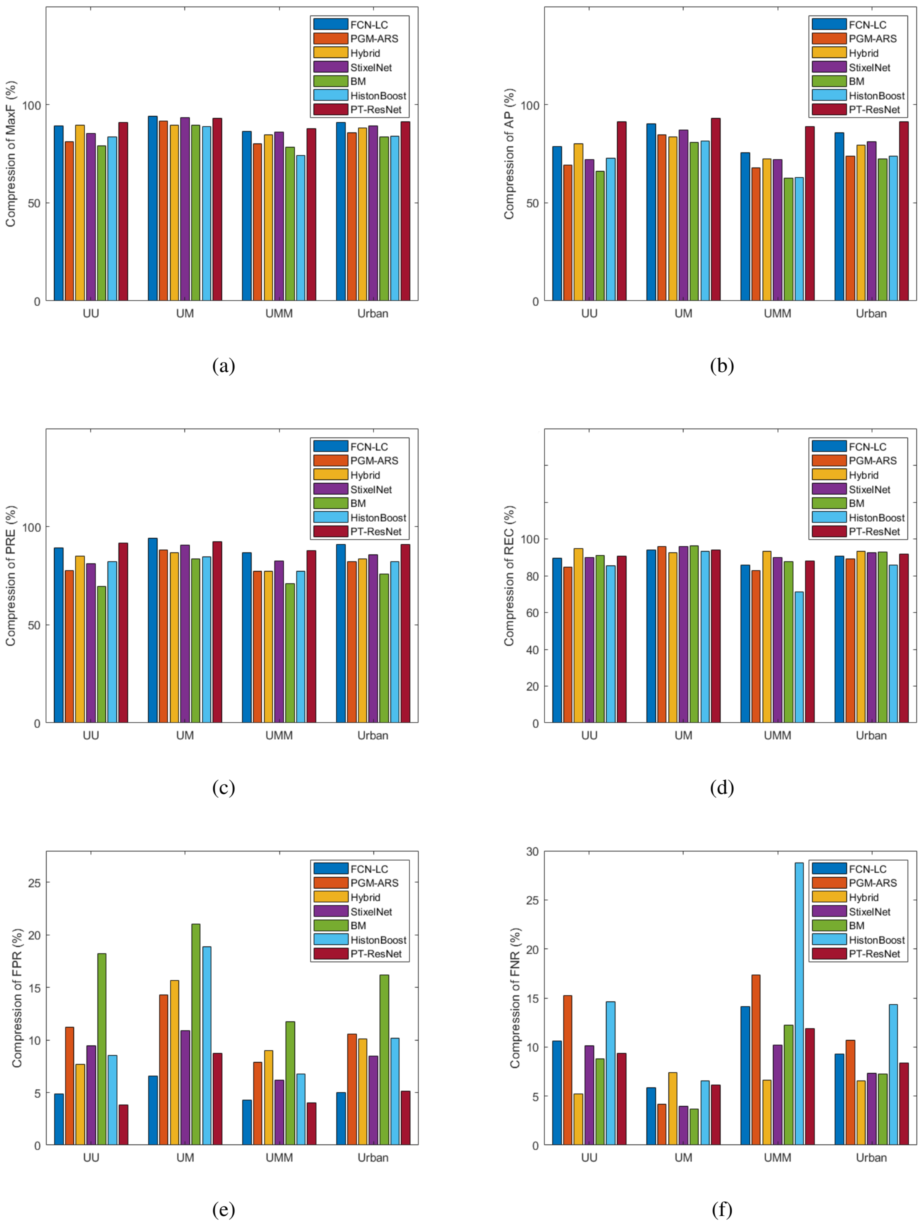
In this paper, we presented a deep neural network for semantic road image segmentation. Since the proposed network focuses entirely on the road surface, the left and right stereo images were processed using our previously published perspective transformation algorithm. This greatly enhanced the similarity of the road surface between the left and right images. The processed stereo images and their corresponding subpixel disparity maps were utilized to create 3D training data. Additionally, we developed our network based on ResNet, a state-of-the-art network with an encoder-decoder structure. According to the evaluation results provided by the KITTI road benchmark, our proposed method outperforms FCN-LC, PGM-ARS, Hybrid, StixelNet, BM and HistonBoost in terms of MaxF and AP, achieving an overall MaxF and AP of 91.19% and 91.21%, respectively. However, the ResNet from DeepLab-v3+ may not be the best network for learning road semantic segmentation from our created 3D training data. Therefore, we plan to train different state-of-the-art networks, such as VGG-16 and VGG-19, and compare the experimental results with what we achieved in this paper.

Acknowledgment

This work was supported by the National Natural Science Foundation of China, under grant No. U1713211, the Research

[1] http://www.cvlibs.net/datasets/kitti/eval_road.php.

Authorized licensed use limited to: Aristotle University of Thessaloniki. Downloaded on September 29,2023 at 11:05:14 UTC from IEEE Xplore. Restrictions apply.



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 5. Performance evaluation. (a) Comparison of MaxF. (b) Comparison of AP. (c) Comparison of PRE. (d) Comparison of REC. (e) Comparison of FPR. (f) Comparison of FNR.

Grant Council of Hong Kong SAR Government, China, under Project No. 11210017, No. 21202816, and the Shenzhen Science, Technology and Innovation Commission (SZSTI) under grant JCYJ20160428154842603, awarded to Prof. Ming Liu.

References

[1] J. A. Brink, R. L. Arenson, T. M. Grist, J. S. Lewin, and D. Enzmann, “Bits and bytes: the future of radiology lies in informatics and information technology,” *European radiology*, vol. 27, no. 9, pp. 3647–3651, 2017.

[2] R. Fan, J. Jiao, H. Ye, Y. Yu, I. Pitas, and M. Liu, “Key ingredients of self-driving cars,” *arXiv:1906.02939*, 2019.

[3] R. Fan and N. Dahnoun, “Real-time stereo vision-based lane detection system,” *Measurement Science and Technology*, vol. 29, no. 7, p. 074005, 2018.

[4] P. Lin, “Why ethics matters for autonomous cars,” in *Autonomous driving*. Springer, Berlin, Heidelberg, 2016, pp. 69–85.

[5] D. Cardinal, “The self-driving industry is finally becoming more realistic,” ExtremeTech, Tech. Rep., Jan. 2019. [Online]. Available: https://www.extremetech.com/extreme/283632-self-driving-industry

[6] U. Ozgunalp, R. Fan, X. Ai, and N. Dahnoun, “Multiple lane detection algorithm based on novel dense vanishing point estimation,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 3, pp. 621–632, 2016.

[7] R. Fan, J. Jiao, J. Pan, H. Huang, S. Shen, and M. Liu, “Real-time dense stereo embedded in a uav for road inspection,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[8] X. Du, M. H. Ang, S. Karaman, and D. Rus, “A general pipeline for 3d detection of vehicles,” in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 3194–3200.

[9] L. Zheng, Y. Zhu, B. Xue, M. Liu, and R. Fan, “Low-cost gps-aided lidar state estimation and map building.”

[10] Y. Zhu, B. Xue, L. Zheng, H. Huang, M. Liu, and R. Fan, “Real-time, environmentally-robust 3d lidar localization.”

Authorized licensed use limited to: Aristotle University of Thessaloniki. Downloaded on September 29,2023 at 11:05:14 UTC from IEEE Xplore. Restrictions apply.

[11] C. Yan, H. Xie, D. Yang, J. Yin, Y. Zhang, and Q. Dai, "Supervised hash coding with deep neural network for environment perception of intelligent vehicles," *IEEE transactions on intelligent transportation systems*, vol. 19, no. 1, pp. 284–295, 2018.

[12] R. Fan, "Real-time computer stereo vision for automotive applications," Ph.D. dissertation, University of Bristol, 2018.

[13] J. Y. Baek, I. V. Chelu, L. Iordache, V. Paunescu, H. Ryu, A. Ghiuta, A. Petreanu, Y. Soh, A. Leica, and B. Jeon, "Scene understanding networks for autonomous driving based on around view monitoring system," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2018, pp. 1074–10 747.

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

[15] C. C. T. Mendes, V. Frémont, and D. F. Wolf, "Exploiting fully convolutional neural networks for fast road detection," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, May 2016, pp. 3174–3179.

[16] M. Passani, J. J. Yebes, and L. M. Bergasa, "Fast pixelwise road inference based on uniformly reweighted belief propagation," in *Proc. IEEE Intelligent Vehicles Symp. (IV)*, Jun. 2015, pp. 519–524.

[17] L. Xiao, R. Wang, B. Dai, Y. Fang, D. Liu, and T. Wu, "Hybrid conditional random field based camera-lidar fusion for road detection," *Information Sciences*, vol. 432, pp. 543–558, 2018.

[18] D. Levi, N. Garnett, E. Fetaya, and I. Herzlyia, "Stixelnet: A deep convolutional network for obstacle detection and road segmentation." in *BMVC*, 2015, pp. 109–1.

[19] B. Wang, V. Frémont, and S. A. R. Florez, "Color-based road detection and its evaluation on the kitti road benchmark," in *IEEE Intelligent Vehicles Symposium (IV 2014)*, 2014, pp. 31–36.

[20] G. B. Vitor, A. C. Victorino, and J. V. Ferreira, "Comprehensive performance analysis of road detection algorithms using the common urban kitti-road benchmark," in *Proc. IEEE Intelligent Vehicles Symp*, Jun. 2014, pp. 19–24.

[21] I. Pitas, *Digital image processing algorithms and applications*. John Wiley & Sons, 2000.

[22] L. 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," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[23] H. Huang, R. Fan, Y. Zhu, M. Liu, and I. Pitas, "A robust pavement mapping system based on normal-constrained stereo visual odometry," 2019.

[24] R. Fan, X. Ai, and N. Dahnoun, "Road surface 3d reconstruction based on dense subpixel disparity map estimation," *IEEE Transactions on Image Processing*, vol. 27, no. 6, pp. 3025–3035, 2018.

[25] J. Chang and Y. Chen, "Pyramid stereo matching network," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition*, Jun. 2018, pp. 5410–5418.

[26] R. Fan and M. Liu, "Road damage detection based on unsupervised disparity map segmentation," *IEEE Transactions on Intelligent Transportation Systems*, 2019.

[27] R. Fan, U. Ozgunalp, B. Hosking, M. Liu, and I. Pitas, "Pothole detection based on disparity transformation and road surface modeling," *IEEE Transactions on Image Processing*, vol. 29, pp. 897–908, 2020.

[28] R. Fan, M. J. Bocus, and N. Dahnoun, "A novel disparity transformation algorithm for road segmentation," *Information Processing Letters*, vol. 140, pp. 18–24, 2018.

[29] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *arXiv preprint arXiv:1802.02611*, 2018.

[30] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[31] J. Fritsch, T. Kühnl, and A. Geiger, "A new performance measure and evaluation benchmark for road detection algorithms," in *Proc. 16th Int. IEEE Conf. Intelligent Transportation Systems (ITSC 2013)*, Oct. 2013, pp. 1693–1700.

Authorized licensed use limited to: Aristotle University of Thessaloniki. Downloaded on September 29,2023 at 11:05:14 UTC from IEEE Xplore. Restrictions apply.