Self-Supervised Learning for Drivable Area and Road Anomaly Segmentation in RGB-D Data

Saritha Kunamalla[1], Gangone Swathi[2], Ramadevi Jaida[3]

[1,2,3]Assistant Professor, Department of ECE, Malla Reddy Engineering College and Management Sciences, Hyderabad, Telangana.

Abstract

Foreground moving object segmentation is a critical challenge in various computer vision applications. Background modeling techniques have made significant progress, but achieving accurate foreground segmentation remains elusive. Most existing algorithms operate exclusively within the color space, making them susceptible to issues like lighting changes, shadows, automatic camera adjustments, and color camouflage. Obtaining large-scale datasets with hand-labeled ground truth is both time-consuming and labor-intensive, rendering these methods challenging to implement in practice. This work presents a self-supervised learning solution for drivable area and road anomaly segmentation, bypassing the need for manual labeling. The proposed method automates the generation of segmentation labels for drivable areas and road anomalies. It leverages RGB-D data to train neural networks for semantic segmentation, using a pipeline known as the Self-Supervised Label Generator (SSLG) to create segmentation labels. The SSLG-generated labels are then employed to train multiple RGB-D data-based semantic segmentation neural networks.

Keywords: Self-Supervised Label Generator, RGB-D, Anomaly Detection, Foreground Segmentation, Computer Vision, Semantic Segmentation.

1. Introduction

Robotic wheelchairs are designed to improve the life quality of the disabled or elderly people by increasing their mobility [1]. To this end, autonomous navigation has been intensively studied and become an essential capability for robotic wheelchairs [2]. The segmentation of drivable areas and road anomalies refers to pixel-wisely identifying the areas and anomalies in images. It is a crucial component for autonomous navigation [3]. Without correctly segmenting drivable areas and road anomalies, robotic wheelchairs could bump or even roll over when passing through road anomalies, which may cause injuries to human riders [4]. In this paper, we define the drivable area as the area where robotic wheelchairs can pass through regardless of their sizes, while the road anomaly is defined as the area with the height larger than 5cm from the surface of the drivable area [5].

The segmentation of drivable areas and road anomalies could be addressed using semantic segmentation techniques [6]. RGB-D cameras, such as Kinect are visual sensors that can stream RGB and depth images at the same time. We use an RGB-D camera for the segmentation of drivable areas and road anomalies in this paper. The reason why we use RGB-D camera is that the depth difference between road anomalies and drivable areas could be useful to distinguish them [7]. Recent development of deep learning techniques has brought significant improvements on the semantic segmentation using RGB-D cameras [8]. To train a deep neural network, we usually need a large-scale dataset with hand-labelled ground truth. However, generating such a dataset is time-consuming and labour-intensive [9]. To provide a solution for the excessive consumption of time and labour for manual labelling, we present a self-supervised [10] approach to segment drivable areas and road anomalies for robotic wheelchairs with an Intel Real sense D415 RGB-D camera.

Rest of the paper is organized as follows: Section 2 details about literature survey, section 3 details about the proposed methodology, section 4 details about the results with discussion, and section 5 concludes article with references.

2. Literature Survey

In [11], proposed an approach for anomaly detection in videosurveillance scenes based on a weekly supervised learning algorithm. Spatio-temporal features are extracted from each surveillance video using a temporal convolutional 3D neural network (TC3D). Then, a novel ranking loss function increases the distance between the classification scores of anomalous and normal videos, reducing the number of false negatives. In [12] introduced FLAGS, the Fused-AI interpretable Anomaly Generation System, and combine both techniques in one methodology to overcome their limitations and optimize them based on limited user feedback. Semantic knowledge is incorporated in a machine learning technique to enhance expressivity. In [13] proposed an approach that combines the advantages and balances the disadvantages of these two methods. An end-to-end network is designed to conduct future frame prediction and reconstruction sequentially. Future frame prediction makes the reconstruction errors large enough to facilitate the identification of abnormal events, while reconstruction helps enhance the predicted future frames from normal events. In [14] aimed to present a real-time vision-based approach that automatically segments the road anomalies from the drivable area. An Intel RealSense D435 depth camera has been employed to capture RGB and depth (RGB-D) images of the road surface. An unsupervised learning method based on diffusion process has been employed to learn the affinity matrix of the RGB-D data and spectral clustering has been applied on the updated affinity matrix to cluster the road images. In [15] proposed, to improve the real-time efficiency of expressway operation monitoring and management, the anomaly detection in intelligent monitoring network of expressway based on edge computing and deep learning is studied. The video data collected by the camera equipment in the intelligent monitoring network structure of the expressway is transmitted to the edge processing server for screening and then sent to the convolutional neural network. The convolutional neural network uses the multi-scale optical flow histogram method to pre-process the video data after the edge calculation to generate the training sample set and send it to the AlexNet model for feature extraction. SVM classifier model is used to train the feature data set and input the features of the test samples into the trained SVM classifier model to realize the anomaly detection in the intelligent monitoring network of expressway.

In [16] proposed a novel method based on the autoencoder. In this method, the latent space of the autoencoder is estimated using a discrete probability model. With the estimated probability model, the anomalous components in the latent space can be well excluded and undesirable reconstruction of the anomalous parts can be avoided. Specifically, we first adopt VQVAE as the reconstruction model to get a discrete latent space of normal samples. In [17] stated a method based on light strip inductive key frame extraction and patchlevel unsupervised network. The light strip inductive feature is designed to simulate the status of train door movement, which contributes to extracting key frames for anomaly detection. An unsupervised network, which is named metro anomaly generative adversarial network (MAGAN) and based on dilate fine-grained generator, multipitch discriminator, and local attention reconstruction loss, is proposed for anomaly classification and localization. In [18] proposed an attention-based model which learns to focus on different parts of a vehicle by conditioning the feature maps on visible key-points. They use triplet embedding to reduce the dimensionality of the features obtained from the ensemble of networks trained using different datasets. To address the problem of anomaly detection, they designed an unsupervised algorithm to detect and localize anomalies in traffic scenes. In [19] presented a deep learning approach for automatic detection and localization of road accidents has been proposed by formulating the problem as anomaly detection.

The method follows one-class classification approach and applies spatiotemporal autoencoder and sequence-to-sequence long short-term memory autoencoder for modelling spatial and temporal representations in the video. The model is executed on a real-world video traffic surveillance datasets and significant results have been achieved both qualitatively and quantitatively. In [20] proposed an exhaustive and systematic literature review of these technologies in RCM that have been published from 2017–2022 by utilizing next-generation sensors, including contact and noncontact measurements. The various methodologies and innovative contributions of the existing literature reviewed in their paper, together with their limitations, promise a futuristic insight for researchers and transport infrastructure owners. The decisive role played by smart sensors and data acquisition platforms, such as smartphones, drones, vehicles integrated with non-intrusive sensors, such as RGB, and thermal cameras, lasers and GPR sensors in the performance of the system are also highlighted.

3. Proposed Methodology

3.1 Dataset Construction

Note that we divide the common road anomalies for robotic wheelchairs into two categories: large road anomalies with a height larger than 15cm from the surface of the drivable area; small road anomalies with a height between 5cm to 15cm from the surface of the drivable area. To the best of our knowledge, this is the first dataset that exhibit common road anomalies for robotic wheelchairs. We use the Real sense camera to collect data involving both large and small road anomalies for the segmentation problem of robotic wheelchairs. Our dataset covers 30 common scenes where robotic wheelchairs usually work (e.g., sidewalks and squares) and 18 different kinds of road anomalies that robotic wheelchairs may encounter in real environments. Fig. 1 shows the number of finely annotated pixels for every kind of road anomaly in our dataset. There are a total of 3896 RGB-D images with hand-labeled ground truth for segmentation in our dataset, which are with the image resolution of 720 × 1280 pixels. It should be noted that our proposed self-supervised approach does not require hand-labeled ground truth. The hand-labeled ground truth is only used for the evaluation of our proposed self-supervised approach. We provide two kinds of depth data in our dataset, the original depth data and the normalized depth data. The normalized depth data is normalized to the range of 0 to 255. Note that the distance measurement range for the Realsense RGB-D camera is up to 10m. Therefore, we remove the pixels with the distance larger than 10m and label them with the zero value. As for the hand-labeled ground truth, we label the unknown area with 0, the drivable area with 1 and road anomalies with 2. The area except the drivable area and road anomalies is defined as the unknown area since it is not clear whether robotic wheelchairs can pass through it such as the area beyond the range of the Realsense camera.

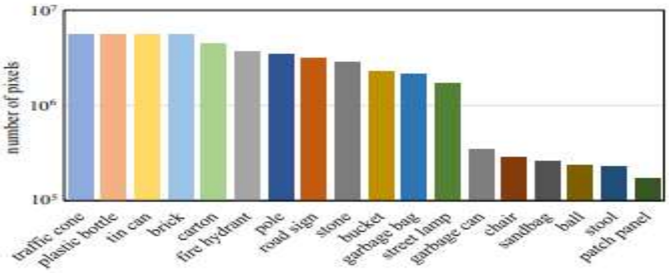


Figure 1: Number of finely annotated pixels (y-axis) and their associated categories (x-axis).

3.2 Self-Supervised Label Generator Our Proposed Self-Supervised Label Generator (SSLG) is designed to generate self-supervised labels of drivable areas and road anomalies automatically. We firstly elaborate the depth processing pipeline inspired by [8]. As derived in [8], for an RGB-D camera consisting of two cameras, the projection of the real-world point P with coordinates of (X, Y, Z) on the image coordinates (U, V) can be computed by (1)– (3):

$$U_l = u_l - u_0 = f\frac{X + b/2}{Y\sin\theta + Z\cos\theta} \qquad (1)$$

$$U_r = u_r - u_0 = f\frac{X - b/2}{Y\sin\theta + Z\cos\theta} \qquad (2)$$

$$V = v - v_0 = f\frac{Y\cos\theta - Z\sin\theta}{Y\sin\theta + Z\cos\theta} \qquad (3)$$

where b is the distance between the optical centers of two cameras; f is the focal length; (u0, v0) is the center of the image plane; ul, ur are the projection of the point P on two cameras, respectively; θ is the pitch angle with respect to the ground plane. Then, the disparity Δ can be calculated by (4):

$$\Delta = u_l - u_r = f\frac{b}{Y\sin\theta + Z\cos\theta} \qquad (4)$$

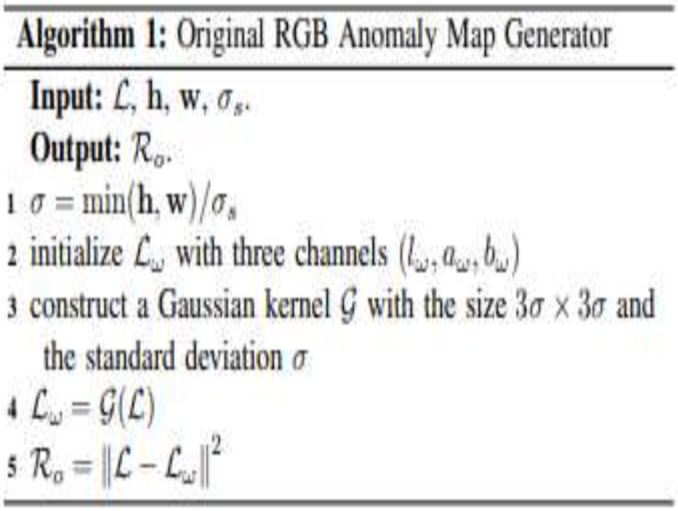
Horizontal planes in the real-world coordinates can be represented by Y = m, which leads to:

$$\Delta\frac{m}{b} = V\cos\theta + f\sin\theta \qquad (5)$$

Similarly, vertical planes in the real-world coordinates can be represented by Z = n, which leads to:

$$\Delta\frac{n}{b} = V\sin\theta + f\cos\theta \qquad (6)$$

Equation (5) and (6) show that horizontal planes and vertical planes in the real-world coordinates can be projected as straight lines in the v-disparity map. Actually, [8] proposed that this conclusion applies to all planes. The intuition behind the depth processing pipeline is that drivable areas can be regarded as planes in most cases and road anomalies can also be regarded as planes approximately. Then, the segmentation problem can be converted into a straight-line extraction problem. The original v-disparity map can be obtained by computing the depth histogram of each row in the depth image. Since the computed v-disparity map often contains much noise, the steerable filter with the second derivatives of Gaussian as the basis function [9] is applied to filter the original v-disparity map. Then, the Hough Transform algorithm [27] is applied to extract straight lines in the filtered v-disparity map. Gao et al. [9] concluded that the drivable area is dominant in v-disparity maps; the straight line with the smallest disparity is the projection of the infinity plane; the remaining straight lines except the two straight lines mentioned above are marked as road anomalies. According to these conclusions, we firstly filter out the straight lines representing road anomalies with a height smaller than 5cm from the surface of the drivable area according to their lengths. Then, it is easy to find that in the filtered v-disparity map, straight line No.1, No.2 and No.3 represent the drivable area, the road anomaly, and the infinity plane, respectively. After that, we extract the drivable area MD and the original depth anomaly map Do according to the straight-line detection results. However, the original depth anomaly map lacks robustness and accuracy because the straight lines representing small road anomalies are too short and easy to be filtered out together with the noise. For instance, there are three road anomalies in the example, but there is only one straight line representing road anomalies in the filtered v-disparity map and thus one road anomaly detected in the original depth anomaly map. The other two small road anomalies (i.e., the brick and the road sign) are filtered out together with the noise. To solve this problem, we utilize the drivable area that we have already generated. We can find that there are some holes inside the drivable area, which contain the missing road anomalies in the original depth anomaly map. Therefore, we extract holes in the drivable area and then combine the hole detection results with the original depth anomaly map to generate the final depth anomaly map Df , which is further normalized to the range [0, 1] (Fig. 3 VI). Although this method will bring some noise to the depth anomaly map, it greatly increases the detection rate of road anomalies to ensure the safety of the riders and we will correct it again with the information of RGB images.



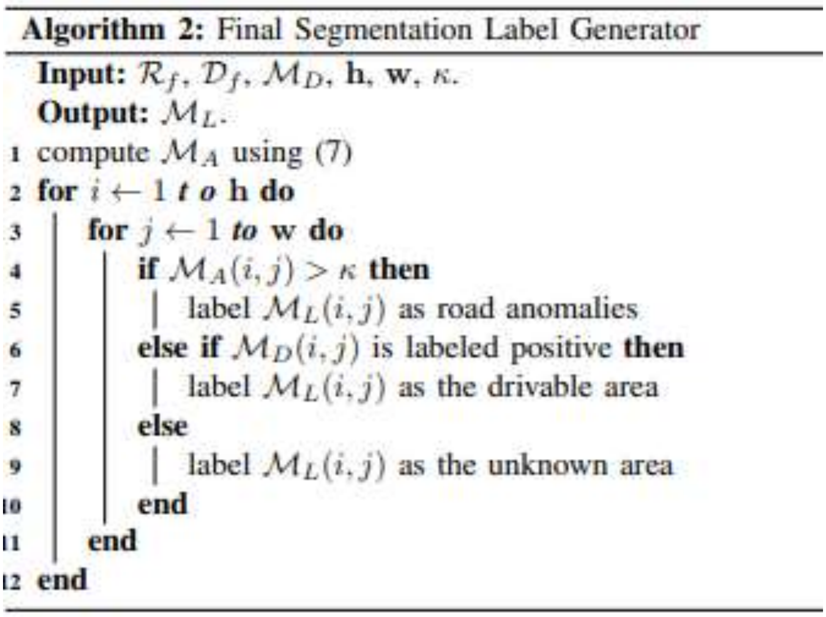
```
Algorithm 1: Original RGB Anomaly Map Generator
Input: $\mathcal{L}$, h, w, $\sigma_s$.
Output: $\mathcal{R}_o$.
$\sigma = \min(\mathbf{h}, \mathbf{w})/\sigma_s$
initialize $\mathcal{L}_\omega$ with three channels $(l_\omega, a_\omega, b_\omega)$
construct a Gaussian kernel $\mathcal{G}$ with the size $3\sigma \times 3\sigma$ and
  the standard deviation $\sigma$
$\mathcal{L}_\omega = \mathcal{G}(\mathcal{L})$
$\mathcal{R}_o = \|\mathcal{L} - \mathcal{L}_\omega\|^2$
```

Now we elaborate the RGB image processing pipeline inspired by [15]. The intuition behind the RGB processing pipeline is that the areas with different colors from surrounding areas are often marked as road anomalies. Based on this principle, we design an original RGB anomaly map generator, which is described in Algorithm 1. Let L denote the image in the Lab color space transformed from the RGB image, and h and w denote the width and the height of the RGB image, respectively.

We generate the original RGB anomaly map Ro by computing the difference between the Lab color vector of each pixel and its Gaussian blurred result. To suppress the pattern of the drivable area, we choose a large filter scale for the Gaussian kernel to blur each channel of the Lab color space. The size of the kernel is 3σ ×3σ with the standard deviation σ and σs is chosen to be 12 to control the strength of weighting. However, the original RGB anomaly map lacks robustness and accuracy because of the interference outside the drivable area. To solve this problem, we utilize the drivable area that we have already generated to filter out the noise outside the drivable area. After normalized to the range [0, 1], the final RGB anomaly map Rf is generated. The last step of our proposed SSLG is to combine two anomaly maps and the drivable area to generate the self-supervised label. We design a final segmentation label generator, which is described in Algorithm 2. As for road anomalies, we firstly generate the final anomaly map MA according to (7):

$$\mathcal{M}_A = \alpha\mathcal{R}_f + (1-\alpha)\mathcal{D}_f \qquad (7)$$

Then, we set a threshold κ and the area in the final anomaly map where the value is greater than κ is marked as road anomalies in the self-supervised label. In our case, we set α to 0.5 and κ to 0.3. The drivable area in the self-supervised label is the same as the drivable area MD, and the rest area except drivable areas and road anomalies marked above is labeled as the unknown area. Finally, the self-supervised label ML is generated, which is used for training RGB-D data-based semantic segmentation neural networks as described in the following sections.

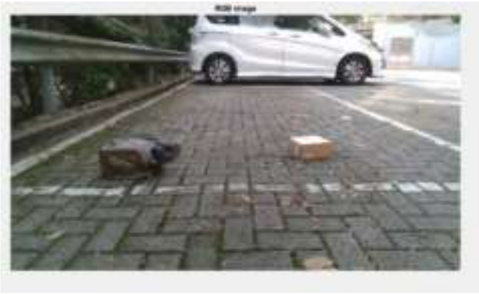


```
Algorithm 2: Final Segmentation Label Generator
Input: $\mathcal{R}_f$, $\mathcal{D}_f$, $\mathcal{M}_D$, h, w, $\kappa$.
Output: $\mathcal{M}_L$.
compute $\mathcal{M}_A$ using (7)
for $i \leftarrow 1$ t o h do
    for $j \leftarrow 1$ to w do
        if $\mathcal{M}_A(i,j) > \kappa$ then
            label $\mathcal{M}_L(i,j)$ as road anomalies
        else if $\mathcal{M}_D(i,j)$ is labeled positive then
            label $\mathcal{M}_L(i,j)$ as the drivable area
        else
            label $\mathcal{M}_L(i,j)$ as the unknown area
        end
    end
end
```

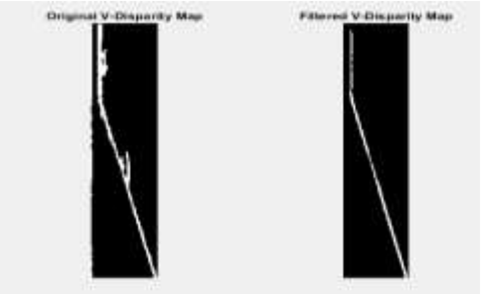
4. Results and Discussions

This section gives the detailed analysis of simulations which are implemented on MATLAB R2016a, and the performance performed using both visual and objective analysis.

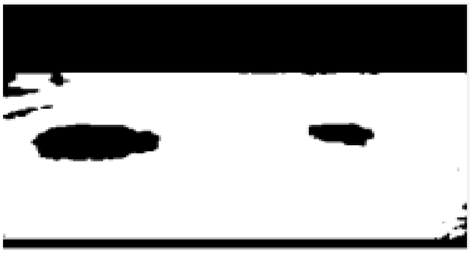
Dataset: The proposed simulations are implemented on the real time RGB-D dataset. RGB-D cameras, such as the Kinect, are visual sensors that are capable of simultaneously streaming RGB and depth pictures. Throughout this research, an RGB-D camera for the segmentation of drivable zones and road abnormalities. Because the depth difference between road anomalies and drivable regions might be valuable in distinguishing between them, we are using an RGB-D camera for this purpose.



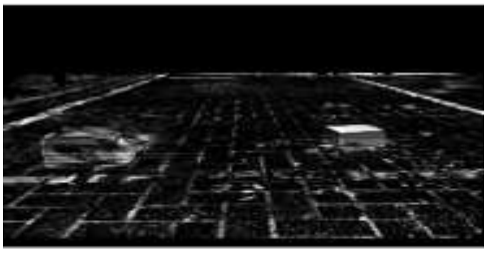
(a)



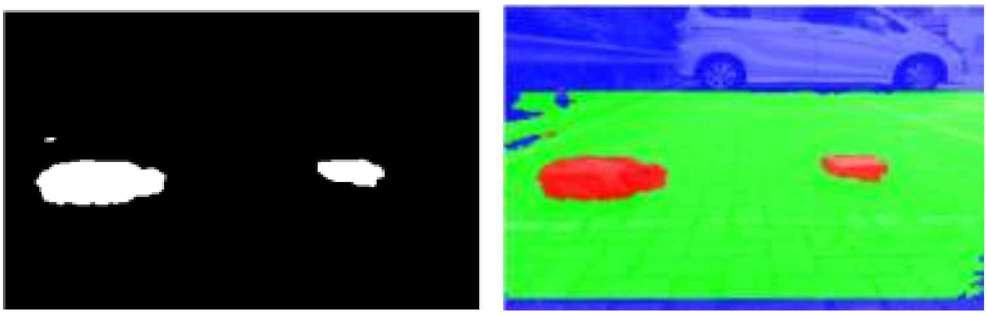
(b)



(c)



(d)



(e) (f)

Figure 2: Simulation results (a) input image, (b) V-Disparity maps, (c) drivable area, (d) RGB anamoly, (e) Final depth anamoly map, (f) self supervised label.

Figure 3 shows the simulational results implemented in MATLAB software, figure 2 (a) shows the original input image. Figure 2 (b) shows the intailly generated v-disparity map and filtered v-disparity map. Figure 2 (c) shows the black and white drivable area with no colour indication and Figure 2 (d) represents the black and white version of RGB anamoly map, respectively. Finally, Figure 2 (e) represents the filtered black and white version of final depth anamoly map and resultant Figure 2 (f) represents the self supervised label, where green colour indicates the drivable area, red colour indicates the anamoly and blue colour indicates the not front view unwanterd area, respectively. Table 1 shows that the performance of the proposed method is improved as compared to the conventional EVD [9], RSL [11], and DSL [13] approches for performance metrics.

Table 1. Performance Comparision

Method	Accuracy	Precision	Recall
EVD [9]	91.785	90.148	90.160
RSL [11]	92.233	92.482	91.622
DSL [13]	94.161	95.005	92.116
Propsoed method	97.651	96.644	92.544

5. Conclusion

In this work, we presented a comprehensive study on the drivable area and road anomaly segmentation problem for robotic wheelchairs. A self-supervised approach was proposed, which contains an automatic labelling pipeline for drivable area and road anomaly segmentation. Experimental results showed that our proposed automatic labelling pipeline achieved an impressive speed-up compared to manual labelling. In addition, our proposed self-supervised approach exhibited more robust and accurate results than the state-of-the-art traditional algorithms as well as the state-of-the-art self-supervised algorithms. In our future work, we plan to investigate our work with planning algorithms for robotic wheelchairs to achieve autonomous navigation.

References

[1]. Z. Zhang, “Microsoft kinect sensor and its effect,” IEEE multimedia, vol. 19, no. 2, pp. 4–10, 2012.

[2]. Y. Sun, M. Liu, and M. Q.-H. Meng, “Active Perception for Foreground Segmentation: An RGB-D Data-Based Background Modeling Method,” IEEE Transactions on Automation Science and Engineering, pp. 1–14, 2019.

[3]. ——, “Motion removal for reliable rgb-d slam in dynamic environments,” Robotics and Autonomous Systems, vol. 108, pp. 115 – 128, 2018.

[4]. ——, “Improving rgb-d slam in dynamic environments: A motion removal approach,” Robotics and Autonomous Systems, vol. 89, pp. 110 – 122, 2017.

[5]. C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, “Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture,” in Asian Conference on Computer Vision. Springer, 2016, pp. 213–228.

[6]. W. Wang and U. Neumann, “Depth-aware cnn for rgb-d segmentation,” in The European Conference on Computer Vision (ECCV), September 2018.

[7]. F. Lateef and Y. Ruichek, “Survey on semantic segmentation using deep learning techniques,” Neurocomputing, 2019.

[8]. R. Labayrade, D. Aubert, and J.-P. Tarel, “Real time obstacle detection in stereovision on non flat road geometry through" v-disparity" representation,” in Intelligent Vehicle Symposium, 2002. IEEE, vol. 2. IEEE, 2002, pp. 646–651.

[9]. Y. Gao, X. Ai, Y. Wang, J. Rarity, and N. Dahnoun, “Uv-disparity based obstacle detection with 3d camera and steerable filter,” in Intelligent Vehicles Symposium (IV), 2011 IEEE. IEEE, 2011, pp. 957–962.

[10]. Y. Cong, J.-J. Peng, J. Sun, L.-L. Zhu, and Y.-D. Tang, “V-disparity based ugv obstacle detection in rough outdoor terrain,” Acta Automatica Sinica, vol. 36, no. 5, pp. 667–673, 2010.

[11]. D. Yiruo, W. Wenjia, and K. Yukihiro, “Complex ground plane detection based on v- disparity map in off-road environment,” in Intelligent Vehicles Symposium (IV), 2013 IEEE. IEEE, 2013, pp. 1137–1142.

[12]. U. Ozgunalp, X. Ai, and N. Dahnoun, “Stereo vision-based road estimation assisted by efficient planar patch calculation,” Signal, Image and Video Processing, vol. 10, no. 6, pp. 1127– 1134, 2016.

[13]. Z. Liu, S. Yu, and N. Zheng, “A co-point mapping-based approach to drivable area detection for self-driving cars,” Engineering, vol. 4, no. 4, pp. 479–490, 2018.

[14]. R. Cong, J. Lei, C. Zhang, Q. Huang, X. Cao, and C. Hou, “Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion,” IEEE Signal Processing Letters, vol. 23, no. 6, pp. 819–823, 2016.

[15]. J. Lou, W. Zhu, H. Wang, and M. Ren, “Small target detection combining regional stability and saliency in a color image,” Multimedia Tools and Applications, vol. 76, no. 13, pp. 14 781– 14 798, 2017.

[16]. H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, “Rgbd salient object detection: a benchmark and algorithms,” in European conference on computer vision. Springer, 2014, pp. 92–109.

[17]. H. Chen, Y. Li, and D. Su, "Multi-modal fusion network with multiscale multi-path and cross-modal interactions for rgb-d salient object detection," Pattern Recognition, vol. 86, pp. 376–385, 2019.

[18]. D. Barnes, W. Maddern, and I. Posner, "Find your own way: Weaklysupervised segmentation of path proposals for urban autonomy," in Robotics and Automation (ICRA), 2017 IEEE International Conference on. IEEE, 2017, pp. 203–210.

[19]. J. Mayr, C. Unger, and F. Tombari, "Self-supervised learning of the drivable area for autonomous vehicles," in 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2018, pp. 362–369.

a. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual worlds as proxy for multi-object tracking analysis," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4340–4349.

[20]. F. Xia, A. R. Zamir, Z. He, A. Sax, J. Malik, and S. Savarese, "Gibson env: Real-world perception for embodied agents," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 9068–9079.