Unifying terrain awareness through real-time semantic segmentation

Kailun Yang[1], Luis M. Bergasa[2], Eduardo Romera[2], Ruiqi Cheng[1], Tianxue Chen[3] and Kaiwei Wang[1]

***Abstract*— Active research on computer vision accelerates the progress in autonomous driving. Following this trend, we aim to leverage the recently emerged methods for Intelligent Vehicles (IV), and transfer them to develop navigation assistive technologies for the Visually Impaired (VI). This topic grows notoriously challenging as it requires to detect a variety of scenes towards higher level of assistance. Computer vision based techniques with monocular detectors or depth sensors sprung up within years of research. These separate approaches achieved remarkable results with relatively low processing time, and improved the mobility of visually impaired people to a large extent. However, running all detectors jointly increases the latency and burdens the computational resources. In this paper, we put forward to seize pixel-wise semantic segmentation to cover the perception needs of navigational assistance in a unified way. This is critical not only for the terrain awareness regarding traversable areas, sidewalks, stairs and water hazards, but also for the avoidance of short-range obstacles, fast-approaching pedestrians and vehicles. At the heart of our proposal is a combination of efficient residual factorized network (ERFNet), pyramid scene parsing network (PSPNet) and 3D point cloud based segmentation. This approach proves to be with qualified accuracy and speed for real-world applications by a comprehensive set of experiments on a wearable navigation system.**

I. INTRODUCTION

Navigational assistance aims to enable visually impaired people to ambulate safely and independently. Challenges stated in this field are frequently related to scene understanding, which are also similar to the problems of autonomous driving. In this regard, the impressive developments of computer vision achieved in Intelligent Vehicles (IV) can be an enormous benefit for the Visually Impaired (VI), supposing crucial prerequisites to enhance vehicular safety as well as pedestrian safety. To extend the coverage of assistance from able-bodied road users to people with visual impairments, many navigational assistive technologies have been developed to accomplish specific goals including avoiding obstacles [1], [2], [3], finding paths [4], [5], [6], [7], locating sidewalks [8], ascending [9] or descending stairs [10], and negotiating water hazards [11].

It is true that each one of these navigational tasks has been tackled well through its respective solutions. However, as the demand of the VI increases [12], this topic grows challenging which requires juggling multiple tasks simultaneously and coordinating all of the perception needs efficiently. Accordingly, the research community has been spurred to integrate different detectors beyond traversability awareness, which is considered as the backbone for any navigational assistive tool [13]. As an illustration, the personal guidance system created in [9] performed two main tasks. It approximately runs the whole floor segmentation at 0.3FPS with additional stair detection iteration time ranging from 50 to 150ms. Even with the high precision in floor detecting and staircase modeling, this approach awaits further optimization to provide assistance at normal walking speed. Multi-threading is an effective way to reduce latency but it increasingly burdens the computational resources. An example is the pair of smart glasses from KR-VISION [14], which detects obstacles, stairs and sidesteps across different processing threads by continuously receiving images from the sensors and multitasking at different frame rates. In a user study of the pRGB-D framework [11], although traversable directions and water puddles were feedback concurrently, demand was revealed for discerning more information of the terrain.

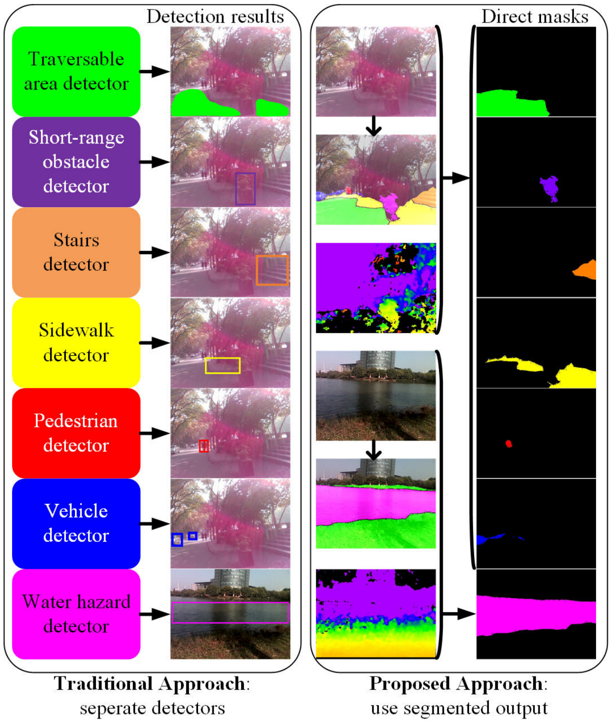

Detection results
Traversable area detector
Short-range obstacle detector
Stairs detector
Sidewalk detector
Pedestrian detector
Vehicle detector
Water hazard detector
Direct masks
Traditional Approach: seperate detectors
Proposed Approach: use segmented output

Fig. 1. Two approaches of perception in navigational assistance for the visually impaired. A different example image was used for water hazards detection, but these images are all captured in real-world scenarios and segmented with the proposed approach.

In the literature, a number of systems rely on sensor fusion to understand more of the surrounding scenes [15]. In another respect, the concept investigated in [16] used a highly integrated radar to warn against collisions with pedestrians and cars, taking into consideration that fast moving objects are response-time critical. However, to the navigation assistance, of even greater concern is the depth data from

[1] Kailun Yang, Ruiqi Cheng and Kaiwei Wang are with College of Optical Science and Engineering, Zhejiang University, Hangzhou, China {elnino, rickycheng, wangkaiwei}@zju.edu.cn; [2] Luis M. Bergasa and Eduardo Romera are with Department of Electronics, University of Alcalá, Madrid, Spain luism.bergasa@uah.es, eduardo.romera@edu.uah.es; [3] Tianxue Chen is with Department of Electrical and Computer Engineering, University of California, Los Angeles, CA, USA tianxuechen@ucla.edu.

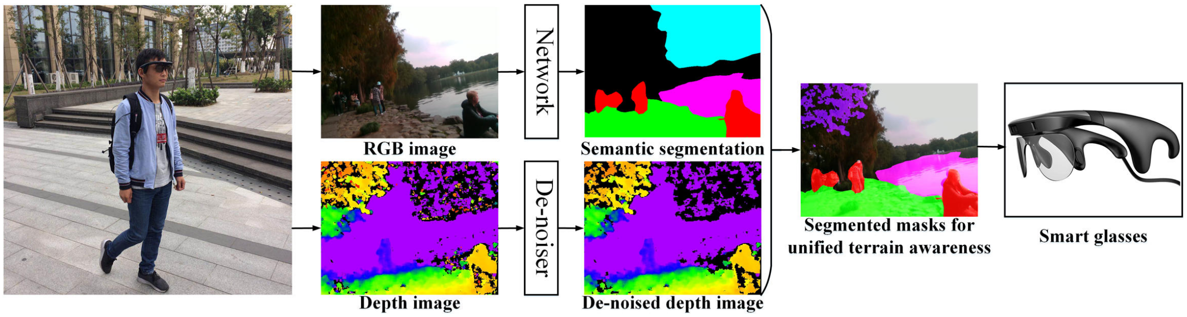


Fig. 2. Overview of the wearable navigation system.

almost all commercial 3D sensors, which suffer from limited depth range and could not maintain the robustness in various environments [13]. Inevitably, approaches with stereo camera or RGB-D sensor generally perform range expansion [2], [17], depth enhancement [5] or depend on both visual and depth information to complement each other [6]. Not to mention the time consumption in these steps, underlying assumptions were frequently made such as the ground plane is the biggest area [1], the area directly in front of the user is accessible [4] and variant versions of Manhattan World [6], [9] or Stixel World assumption [3], [7]. These factors all limit the flexibility in navigational assistive applications.

However, unlike traditional approaches mentioned above, convolution neural networks, learn and discriminate between different features directly from the input data using a deeper abstraction of representation layers. Namely, recent advances in deep learning have achieved break-through results in most vision-based tasks including semantic segmentation, which is to partition an image into several coherent semantically meaningful parts. As depicted in Fig. 1, since traditional approaches detect different targets independently [18], the assistance for the VI are treated separately. Naturally, it is beneficial to provide terrain awareness in a unified way, because it allows to solve many tasks at once and exploit their inter-relations and contexts. Semantic segmentation targets at solving exactly this problem. It classifies a wide variety of scene classes directly leading to pixel-wise understanding, which supposes a very rich source of processed information for higher-level navigational assistance.

Up until very recently, pixel-wise semantic segmentation was not usable in terms of speed. However, a fraction of networks has focused on the efficiency by proposing architectures that could reach near real-time segmentation [19], [20]. These advances have made possible the utilization of full scene segmentation in time-critical cases like blind assistance. Nonetheless, to the best of our knowledge, no previous work has developed real-time semantic segmentation to assist visually impaired pedestrians. Based on this notion, instead of simply identifying the most traversable direction [11], we make an pioneering attempt to provide terrain awareness in a unified way. In this paper, we extend our previous efficient residual factorized network (ERFNet) [20] by combining a pyramid scene parsing network (PSPNet) [21] to respond to the surges in demand. Additionally, a set of fast depth post-processing are implemented to enhance collision avoidance. The main contributions of our work are threefold:

- A unification of terrain awareness regarding traversable areas, obstacles, sidewalks, stairs, water hazards, pedestrians and vehicles.
- A real-time semantic segmentation network to learn both global scene contexts and local textures without imposing any assumptions.
- A real-world navigational assistance framework on a wearable prototype for visually impaired individuals.

The remainder of this paper is structured as follows. In Section II, the framework is elaborated in terms of the wearable assistance system, the semantic segmentation architecture and the implementation details. In Section III, the approach is evaluated and discussed as for real-time and real-world performance. In Section IV, relevant conclusions are drawn and future works are expected.

II. APPROACH

A. Wearable navigation system

In this work, the main motivation is to design a prototype which should be wearable without hurting the self-esteem of visually impaired people. With this target in mind, we follow the trend of using head-mounted glasses to acquire environment information and interact with visually impaired users. As worn by the user in Fig. 2, the system is composed of a pair of smart glasses and a laptop in the backpack. The pair of smart glasses named Intoer, commercially available at [14], is comprised of a RGB-D sensor of RealSense R200 [22] and a set of bone conducting earphones. We utilize a laptop with Core i7-7700HQ processor and GTX 1050Ti as the computing platform, which could be easily carried in a backpack and is robust enough to operate in rough terrain.

This pair of glasses captures real-time RGB-D streams and transfers them to the processor, while the RGB images are fed to the network for semantic segmentation. As for the depth images, which are acquired with the combination of active speckle projecting and passive stereo matching, they are preprocessed in the first place. To enforce the stereo matching algorithm to deliver dense maps, we use a different preset configuration with respect to the original depth image of RealSense by controlling how aggressive the algorithm is at discarding matched pixels. After that, the depth images are de-noised by eliminating small segments which was previously presented in [5]. The dense depth image with noise reduction leads to robust segmentation of short-range obstacles when using the semantic segmentation output as

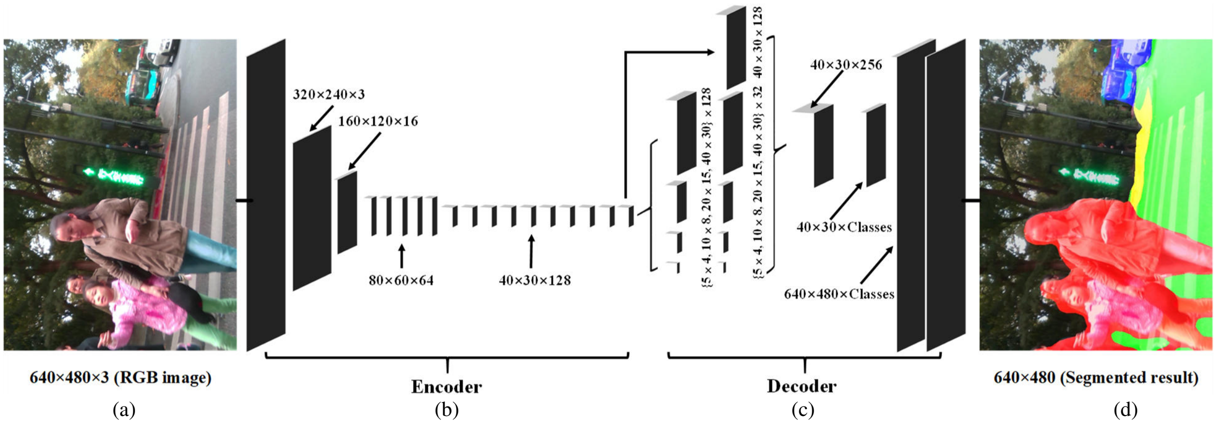


Fig. 3. The proposed architecture. From left to right: (a) Input, (b) Encoder, (c) Decoder, (d) Prediction.

the base for higher-level assistance. As far as the feedback is concerned, the bone conducting earphones transfer the detection results to the VI for both terrain awareness and collision avoidance. This is important as visually impaired people need to continue hearing environmental sounds and the bone conducting interface allows them to hear a layer of augmented acoustic reality that is superimposed on the environmental sounds.

B. Semantic segmentation architecture

In order to leverage the success of segmenting a variety of scenes and maintaining the efficiency, we design the architecture according to the SegNet-based encoder-decoder architecture like ENet [19] and our previous ERFNet [20]. In FCN-like architectures, feature maps from different layers need to be fused to generate a fine-grain output. As indicated in Fig. 3, our approach contrarily uses a more sequential architecture based on an encoder producing down-sampled feature maps and a subsequent decoder that up-samples the feature maps to match input resolution. Table I gives a detailed description of the integral architecture, where residual layers were stacked in the encoder. Generally, the residual layer adopted in state-of-art networks [19], [23] has two instances: the bottleneck version and the non-bottleneck design. In our previous work [20], "Non-bottleneck-1D" (non-bt-1D) was proposed, which is a redesign of the residual layer to leverage the efficiency of the bottleneck and the learning capacity of non-bottleneck in a judicious trade-off way by using 1D factorizations of the convolutional kernels. Thereby, it enables an efficient use of minimized amount of residual layers to extract feature maps and achieve semantic segmentation in real time.

TABLE I
LAYER DISPOSAL OF OUR PROPOSED NETWORK. "OUT-F": NUMBER OF FEATURE MAPS AT LAYER'S OUTPUT, "OUT-RES": OUTPUT RESOLUTION FOR INPUT SIZE OF 640×480.

	Layer	Type	Out-F	Out-Res
ENCODER	0	Scaling 640×480	3	320×240
	1	Down-sampler block	16	160×120
	2	Down-sampler block	64	80×60
	3-7	5×Non-bt-1D	64	80×60
	8	Down-sampler block	128	40×30
	9	Non-bt-1D (dilated 2)	128	40×30
	10	Non-bt-1D (dilated 4)	128	40×30
	11	Non-bt-1D (dilated 8)	128	40×30
	12	Non-bt-1D (dilated 16)	128	40×30
	13	Non-bt-1D (dilated 2)	128	40×30
	14	Non-bt-1D (dilated 4)	128	40×30
	15	Non-bt-1D (dilated 8)	128	40×30
	16	Non-bt-1D (dilated 2)	128	40×30
DECODER	17a	Original feature map	128	40×30
	17b	Pooling and convolution	32	40×30
	17c	Pooling and convolution	32	20×15
	17d	Pooling and convolution	32	10×8
	17e	Pooling and convolution	32	5×4
	17	Up-sampler and concatenation	256	40×30
	18	Convolution	C	40×30
	19	Up-sampler	C	640×480

However, for the terrain awareness in intelligent assistance, we attach a different decoder with respect to the previous work. This key modification aims to collect more contextual information while minimizing the sacrifice of learning textures. Global context information is of cardinal significance for terrain awareness in order to prevent generating confusing feedback. To detail this, if the network mis-predicts a safe path in front of a lake, the VI would be left vulnerable in the dynamic environments. This kind of problem could be remedied by exploiting more context and learning more relationship between categories. With this target in mind, we reconstruct the decoder architecture. In this work, the decoder architecture follows the pyramid pooling module as introduced by PSPNet [21]. This module is leveraged to harvest different sub-region representations, followed by up-sampling and concatenation layers to form the final feature representation. As a result, it carries both local and global context information from the diverse pooled representations at different locations. Since it fuses features under a group of different pyramid levels, the output of different levels in this pyramid pooling module contains the feature map from the encoder with varied sizes. To maintain the weight of global feature, we append a convolution layer after each pyramid level to reduce the dimension of context representation to $1/N$ of the original one if the level size of pyramid is N. As for the situation in Fig. 3c, the level size N equals to 4 and we decrease the number of feature maps from 128 to 32. Subsequently, the low-dimension feature maps are directly up-sampled to obtain the same size features as the original feature map through bilinear interpolation. Fig. 3 contains a depiction of the feature maps generated by each of the block in our architecture, from the RGB input to the per-pixel class probabilities and final dense predictions.

C. Implementation details

Smart glasses. We start a stream of 640×480 RGB image, a stream of 320×240 infrared stereo pair which produces a stream of 320×240 depth image. The depth information are projected to the field of view of color camera so as to acquire a synchronized 640×480 depth stream. To achieve high environmental adaptability, the automatic exposure and gain control are enabled. Most of the depth control thresholds are in the loosest setting while only the left-right consistency constraint is adjusted to 30. For the short-range obstacle avoidance, 5m is set as the threshold to segment directly at pixel level if not classified as traversable area, stairs, water, pedestrian or car.

Dataset. The challenging ADE20K dataset [24] is chosen as it covers both indoor and outdoor scenarios. Also, this dataset contains the classes of stairs and water areas, which are very important scenes for the navigation assistance. To enrich the training dataset, we add the images which have the classes of sky, floor, road, grass, sidewalk, ground, water and stairs from PASCAL-Context dataset [25] and COCO-Stuff 10K dataset [26]. Hence, the training involves 37075 images, within which 20210 images are from ADE20K, 8733 images are from PASCAL-Context and the remaining 8132 images come from COCO-Stuff. In addition, we have 2000 images from ADE20K for validation. To provide awareness regarding the scenes that visually impaired people care the most during navigation, we only use the most frequent 22 classes of scenes or objects for training. Additionally, we merge the water, sea, river, pool and lake into a class of water hazards. In a similar way, the stairs, stairway, staircase are merged into a class of stairs.

Data augmentations. To robustify the model against the varied types of images from real world, we perform a group of data augmentations. Firstly, random cropping and random scaling are jointly used to resize the cropped regions into 320×240 input images. Secondly, a random rotation ranges from -20^{o} to 20^{o} is implemented without cropping. This intuition comes from that during navigation, the orientation of the smart glasses would be constantly changing and the images rotate. Thirdly, color jittering in terms of brightness, saturation, contrast and hue are applied. Jittering factors regarding brightness, saturation, and contrast here are chosen uniformly from 0.8 to 1.2. Hue augmentation is performed by adding a value between -0.2 and 0.2 to the hue value channel of the HSV representation.

Training setup. Our model is trained using the Adam optimization of stochastic gradient descent. Training is operated with a batch size of 12, momentum of 0.9, weight decay of 2×10^{-4}, and we start with a original learning rate of 5×10^{-5} and decrease the learning rate exponentially across epochs. Following the scheme customized in [19], the weights are determined as w_{class} = 1/ln(c+p_{class}), while c is set to 1.001 to enforce the model to learn more information of the less frequent of classes in the dataset. We first adapt the encoder's last layers to produce a single classification output by adding extra pooling layers and a fully connected layer and finally train the modified encoder on ImageNet. After that, the extra layers are removed and the decoder is appended to train the full network. With this setup, the training reaches convergence when cross-entropy loss value is used as the training criterion.

III. EXPERIMENTS AND DISCUSSION

Experiment setup. The experiments are performed with the wearable navigation system in public spaces around Westlake, the Zijingang Campus and the Yuquan Campus at Zhejiang University in Hangzhou, the Polytechnic School at University of Alcalá in Madrid as well as Venice Beach and University of California in Los Angeles. This allows us to evaluate not only on large-scale scene parsing dataset like ADE20K [24], but also on real-world egocentric images, using Intersection-over-Union (IoU) and Pixel-wise Accuracy (P-A) metrics.

Real-time performance. The total computation time of a single frame is 16ms, while the image acquisition and preprocessing from the smart glasses take 3ms, and the time cost for the semantic segmentation is 13ms. In this sense, the computation cost is saved to maintain a reasonably qualified refresh-rate of 62.5FPS on a processor with a single GPU GTX 1050Ti. This inference time demonstrates that it is able to run our approach in real time, while allowing additional time for auditory [1], [5], [11] or tactile feedback [2]. Our ERF-PSPNet inherits the encoder design but implements a quite efficient version of decoder. Thereby, the speed is even slightly faster than our previous approach with ERFNet, which runs at 55.6FPS on the same processor. Additionally, on a embedded GPU Tegra TX1 (Jetson TX1) that enables higher portability while consuming less than 10 Watts at full load, our approach achieves approximately 22.0FPS.

Segmentation accuracy. The accuracy of our approach is evaluated on both the challenging ADE20K dataset and our real-world dataset. This terrain awareness dataset is publicly available at [27], which contains 120 images with fine annotations of important classes for navigation assistance including ground, sidewalk, stairs, water hazards, person and cars. After merging some classes towards better assistance, we evaluate our approach by comparing the proposed architecture ERF-PSPNet, an existing deep convolutional neural network ENet [19] and our previous work ERFNet [20] on the ADE20K validation dataset. Here, the accuracy results are reported using the commonly adopted Intersection-over-Union (IoU) metric. From Table II(a), it could be told that the accuracy of most classes obtained with the proposed ERF-PSPNet exceeds the state-of-the-art architectures that are also designed for real-time applications. Our architecture builds upon previous work but has the ability to collect more contextual information without major sacrifice of learning from textures. As a result, for large-scale scene parsing task that requires greater strength to gather diverse levels of context, only the accuracy of sky and person are slightly lower than ERFNet, which arguably could support more reliable upper-level assistance.

TABLE II

ACCURACY ANALYSIS.

Architecture	Sky	Floor	Road	Grass	Sidewalk	Ground	Person	Car	Water	Stairs	Mean IoU
ENet [19]	89.7%	72.4%	69.4%	56.5%	38.2%	75.0%	26.7%	64.8%	67.3%	23.7%	58.4%
ERFNet [20]	**93.2%**	77.3%	71.1%	64.5%	46.1%	76.3%	**39.7%**	70.1%	67.9%	24.1%	63.1%
ERF-PSPNet	93.0%	**78.7%**	**73.8%**	**68.7%**	**51.6%**	**76.8%**	39.4%	**70.4%**	**77.0%**	**30.8%**	**66.0%**

(a) On ADE20K dataset [24] using Intersection-over-Union (IoU).

Approach	IoU	Pixel-wise Accuracy (P-A)					
		In total	With Depth	Within 2m	2-3m	3-5m	5-10m
3D-RANSAC-F [1]	50.1%	67.2%	73.3%	53.9%	91.8%	85.2%	61.7%
ENet [19]	62.4%	85.2%	88.4%	79.9%	84.3%	89.7%	93.1%
ERF-PSPNet	**82.1%**	**93.1%**	**95.9%**	**96.0%**	**96.3%**	**96.2%**	**96.0%**

(b) On real-world dataset [27] in terms of traversability awareness.
Floor, road, grass, sidewalk and ground are merged into a class of traversable area.
"With Depth": Only the pixels with valid depth information are evaluated.

Accuracy term	Sky	Traversable area	Ground	Sidewalk	Stairs	Water	Person	Car
IoU	88.0%	82.1%	72.7%	55.5%	67.0%	69.1%	66.8%	67.4%
Pixel-wise Accuracy	95.3%	93.1%	81.2%	93.1%	90.1%	86.3%	90.8%	93.1%
With Depth	N/A	95.9%	84.9%	93.1%	90.8%	89.8%	90.4%	92.7%
Within 2m	N/A	96.0%	76.9%	95.0%	91.9%	96.2%	97.7%	94.3%
2-3m	N/A	96.3%	81.7%	96.5%	91.9%	82.3%	93.7%	95.2%
3-5m	N/A	96.2%	87.4%	94.5%	89.4%	76.9%	93.6%	90.8%
5-10m	N/A	96.0%	86.6%	93.6%	93.1%	84.3%	87.4%	91.4%

(c) ERF-PSPNet on real-world dataset [27] in terms of terrain awareness.

To analyze the major concern of detection performance for real-world assistance, we collect results over several depth ranges: within 2m, 2-3m, 3-5m and 5-10m. In navigational assistance, 2m is the general distance for avoiding static obstacles while the warning distance should be longer when a moving object approaches, e.g. 3m for pedestrians and 5m for cars. In addition, the short-range of ground area detection helps to determine the most walkable direction, while superior path planning could be supported by longer traversability awareness, e.g. 5-10m. Table II(b) shows both the IoU and Pixel-wise Accuracy (P-A) of traversability awareness, which is the core task of navigational assistance. Here, the traversable areas involve the ground, floor, road, grass and sidewalk. We compare the traversable area detection of our ERF-PSPNet to a state-of-the-art architecture ENet and a depth based segmentation approach 3D-RANSAC-F [1], which estimates the ground plane based on RANSAC and filtering techniques by using the dense disparity map. As the depth information of the ground area may be noisy and missing in dynamic environments, we implemented a RGB image guided filter [5] to fill holes before detection. Thereupon, the traditional 3D-RANSAC-F achieves decent accuracy ranging from 2m to 5m and it excels ENet from 2m to 3m as the depth map within this range is quite dense thanks to the active stereo design. Still, our ERF-PSPNet outperforms ENet and 3D-RANSAC-F in both ranges. As far as terrain awareness is concerned, even if the IoU is not very high, the segmentation results are still of great use. For the VI, it is preferred to know that there are stairs or there is an approaching pedestrian in some direction even if the shape is not exactly accurate. Also, it is observed in Table II(c) that most of the Pixel-level Accuracy (P-A) within different ranges are over 90%, which reveals the capacity of our approach for the unification of these detection tasks. Fig. 4 exhibits a group of pixel-wise results generated by our ERF-PSPNet, ENet, 3D-RANSAC-F, and Stixel-level segmentation rendered by a procedure FreeSpaceParse [7]. On the one hand, our approach yields longer and more consistent segmentation which will definitely benefit the traversable area detection. On the other hand, it shows very promising results for providing the terrain awareness within this unified framework.

IV. CONCLUSIONS

Navigational assistance for the Visually Impaired (VI) is undergoing a monumental boom thanks to the developments of Intelligent Vehicles (IV) and computer vision. However, monocular detectors or depth sensors are generally applied in separate tasks. In this paper, we derive achievability results for these perception tasks by utilizing real-time semantic segmentation. The proposed framework, based on deep convolutional neural network and depth segmentation, not only benefits the essential traversability awareness at both short and long ranges, but also covers the needs of terrain awareness in a unified manner. In the future, we aim to continuously improve our navigation assistive approach, precisely to incorporate polarization imaging and user studies.

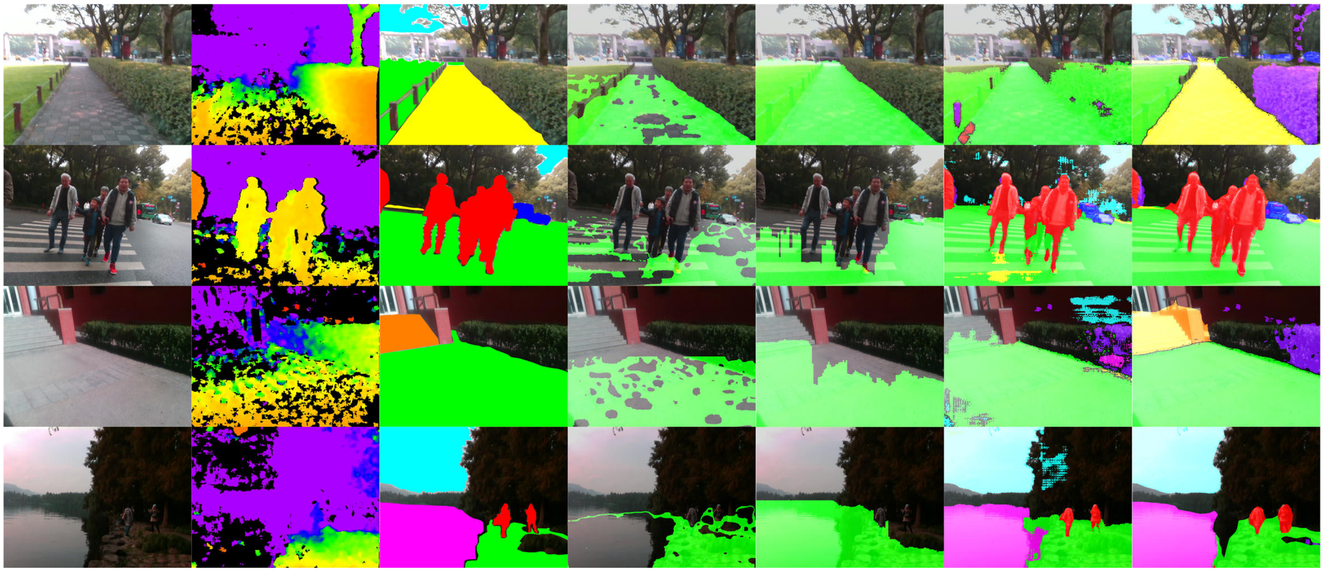
ACKNOWLEDGEMENT

This work has been partially funded by the Zhejiang Provincial Public Fund through the project of visual assistance technology for the blind based on 3D terrain sensor (No. 2016C33136) and cofunded by State Key Laboratory of Modern Optical Instrumentation.

This work has been partially funded by the Spanish MINECO/FEDER through the SmartElderlyCar project (TRA2015-70501-C2-1-R), the DGT through the SERMON project (SPIP2017-02305), and from the RoboCity2030-III-CM project (Robótica aplicada a la mejora de la calidad de vida de los ciudadanos, fase III; S2013/MIT-2748), funded by Programas de actividades I+D (CAM) and cofunded by EU Structural Funds.

REFERENCES

[1] A. Rodríguez, J. J. Yebes, P. F. Alcantarilla and L. M. Bergasa, J. Almazán and A. Cela, "Assisting the visually impaired: obstacle detection and warning system by acoustic feedback," *Sensors*, 2012, 12(12), pp. 17476-17496.



(a) RGB image (b) Depth image (c) Annotation (d) 3D-RANSAC-F (e) FreeSpaceParse (f) ENet (g) Our approach

Fig. 4. Qualitative examples of the segmentation on real-world images produced by our approach compared with ground-truth annotation, 3D-RANSAC-F [1], FreeSpaceParse [7] and ENet [19]. From left to right: (a) RGB image, (b) Depth image, (c) Annotation, (d) 3D-RANSAC-F, (e) FreeSpaceParse, (f) ENet, (g) Our approach.

[2] K. Yang, K. Wang, X. Zhao, R. Cheng, J. Bai, Y. Yang and D. Liu, "IR stereo RealSense: Decreasing minimum range of navigational assistance for visually impaired individuals," *Journal of Ambient Intelligence and Smart Environments*, 2017, 9(6), pp. 743-755.

[3] M. Martinez, A. Roitberg, D. K. B. Schauerte and R. Stiefelhagen, "Using Technology Developed for Autonomous Cars to Help Navigate Blind People," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1424-1432.

[4] D. Koester, B. Schauerte and R. Stiefelhagen, "Accessible section detection for visual guidance," in *Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1-6.

[5] K. Yang, K. Wang, W. Hu and J. Bai, "Expanding the Detection of Traversable Area with RealSense for the Visually Impaired," *Sensors*, 2016, 16(11), 1954.

[6] A. Aladren, G. López-Nicolás, L. Puig and J. J. Guerrero, "Navigation assistance for the visually impaired using RGB-D sensor with range expansion," *IEEE Systems Journal*, 2016, 10(3), pp. 922-932.

[7] H. C. Wang, R. K. Katzschmann, S. Teng, B. Araki, L. Giarré and D. Rus, "Enabling independent navigation for visually impaired people through a wearable vision-based feedback system," in *Robotics and Automation (ICRA), 2017 IEEE International Conference on*. IEEE, 2017, pp. 6533-6540.

[8] F. Ahmed and M. Yeasin, "Optimization and evaluation of deep architectures for ambient awareness on a sidewalk," in *Neural Networks (IJCNN), 2017 International Joint Conference on*. IEEE, 2017, pp. 2692-2697.

[9] J. J. Guerrero, A. Perez-Yus, D. Gutierrez-Gomez, A. Rituerto and G. Lopez-Nicolas, "Human navigation assistance with a RGB-D sensor," 2015, pp. 285-312.

[10] C. Stahlschmidt, S. von Camen, A. Gavriilidis and A. Kummert, "Descending step classification using time-of-flight sensor data," in *Intelligent Vehicles Symposium (IV), 2015 IEEE*. IEEE, 2015, pp. 362-367.

[11] K. Yang, K. Wang, R. Cheng, W. Hu, X. Huang and J. Bai, "Detecting Traversable Area and Water Hazards for the Visually Impaired with a pRGB-D Sensor," *Sensors*, 2017, 17(8), 1890.

[12] S. Wang and L. Yu, "Everyday information behavior of the visually impaired in China," *Information Research*, 2017, 22(1).

[13] K. Saleh, R. A. Zeineldin, M. Hossny, S. Nahavandi and N. A. El-Fishawy, "Navigational Path Detection for the Visually Impaired using Fully Convolutional Networks," in *Systems, Man, and Cybernetics (SMC), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1399-1404..

[14] KR-VISION Technology, "To tackle the challenges for the visually impaired," http://krvision.cn/, 2016.

[15] J. R. Rizzo, Y. Pan, T. Hudson, E. K. Wong and Y. Fang, "Sensor fusion for ecologically valid obstacle identification: Building a comprehensive assistive technology platform for the visually impaired," in *Modeling, Simulation, and Applied Optimization (ICMSAO), 2017 7th International Conference on*. IEEE, 2017, pp. 1-5.

[16] P. Kwiatkowski, T. Jaeschke, D. Starke, L. Piotrowsky, H. Deis and N. Pohl, "A concept study for a radar-based navigation device with sector scan antenna for visually impaired people," in *Microwave Bio Conference (IMBIOC), 2017 First IEEE MTT-S International*, IEEE, 2017, pp. 1-4.

[17] K. Yang, K. Wang, H. Chen and J. Bai, "Reducing the minimum range of a RGB-depth sensor to aid navigation in visually impaired individuals," *Applied Optics*, 2018, 57(11), 2809-2819.

[18] E. Romera, L. M. Bergasa and R. Arroyo, "Can we unify monocular detectors for autonomous driving by using the pixel-wise semantic segmentation of CNNs?" *arXiv preprint arXiv:1607.00971*, 2016.

[19] A. Paszke, A. Chaurasia, S. Kim and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," *arXiv preprint arXiv:1606.02147*, 2016.

[20] E. Romera, J. Alvarez, L. M. Bergasa and R. Arroyo, "ERFNet: Efficient Residual Factorized ConvNet for Real-Time Semantic Segmentation," *IEEE Transactions on Intelligent Transportation Systems*, 2017, 19(1), 263-272.

[21] H. Zhao, J. Shi, X. Qi, X. Wang and J. Jia, "Pyramid scene parsing network," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2881-2890.

[22] L. Keselman, J. I. Woodfill, A. Grunnet-Jepsen and A. Bhowmik, "Intel (R) RealSense (TM) Stereoscopic Depth Cameras," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*. IEEE, 2017, pp. 1267-1276.

[23] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[24] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso and A. Torralba, "Semantic understanding of scenes through the ADE20K dataset," *arXiv preprint arXiv:1608.05442*, 2016.

[25] R. Mottaghi, X. Chen, X. Liu, N. G. Cho, S. W. Lee, S. Fidler, et al., "The role of context for object detection and semantic segmentation in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 891-898.

[26] H. Caesar, J. Uijlings and V. Ferrari, "COCO-Stuff: Thing and Stuff Classes in Context," *arXiv preprint arXiv:1612.03716*, 2016.

[27] Kaiwei Wang Team, "Terrain Awareness Dataset," http://wangkaiwei.org/projecteg.html, 2017.