Semantic perception of curbs beyond traversability for real-world navigation assistance systems

Kailun Yang[1], Luis M. Bergasa[2], Eduardo Romera[2], Dongming Sun[3], Kaiwei Wang[1] and Rafael Barea[2]

***Abstract*— Intelligent Vehicles (IV) and navigational assistance for the Visually Impaired (VI) are becoming highly coupled, both fulfilling safety-critical tasks towards the utopia of all traffic participants. In this paper, the main purpose is to leverage recently emerged methods for self-driving technology, and transfer them to augment perception and aid navigation in ambient assisted living. More precisely, we put forward to seize pixel-wise semantic segmentation to support curb negotiation and traversability awareness, along the pathway of visually impaired individuals. At the crux of our perception unification framework is an effort to attain efficient understanding by proposing a deep architecture built on residual factorized convolution and pyramidical representation. A comprehensive set of experiments demonstrates the accurate scene parsing results with promise of real-time inference speed. Crucially, real-world performance over state-of-art approaches qualifies the proposed framework for assistance when deployed to two wearable navigation systems, including a pair of commercial smart glasses and a prototype of customized device.**

I. INTRODUCTION

Autonomous driving of Intelligent Vehicles (IV) and ambient assisted navigation of pedestrians are becoming tightly intertwined [1] to optimize traffic flow. These two fields confront the fundamental issues, precisely vehicular and pedestrian safety towards the utopia of all traffic participants. To this end, there is a necessity to expand the coverage of assistance from drivers to pedestrians, especially those with visual impairments, who are the most vulnerable road users.

Inspired by the synergy that semantic scene understanding is crucial to enable safe vehicle navigation as well as to enhance mobility of the Visually Impaired (VI) [2], a hotspot has emerged over the past few years. It seeks to leverage the striking advances in autonomous driving, and transfer them to develop navigational assistive technologies [3][4] based on such cross-domain transfer. Along this line, a large body of researches focused on traversability perception [5][6] that constitutes the backbone of any personal guidance system. Beyond the proof-of-concepts established in these researches, the community has also been motivated to provide assistive awareness by integrating stairs detection and water hazards detection [4] at the basis of traversability analysis. In spite of the impressive strides towards higher independence of the VI, curbs negotiation represents a challenging and so far largely unexplored task.

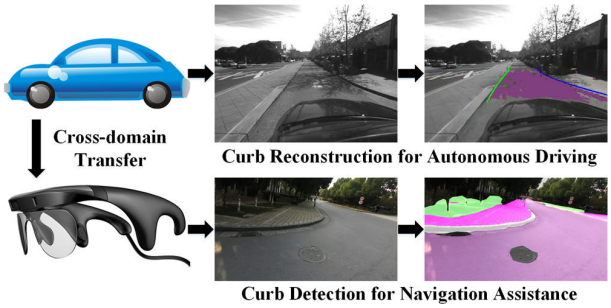


Fig. 1. Along the IV-VI cross-domain research line, this paper proposes to aid in semantic perception of curb for the Visually Impaired (VI), in view of the insufficient reconstruction approach for Intelligent Vehicles (IV).

As revealed by the field test of a previous work [4], navigation on sidewalks and other walkways comprises a major ingredient of the VI's independent living, while suffering a lot from negotiating curbs. More precisely, hazardous curbs along these pathways, pose enormous threats for them during everyday self-navigation, especially within metropolitan areas. In order to close the gap and further enhance safety beyond traversability awareness, we derive insight from the field of IV given the following facts:

- Full pixel-wise semantic segmentation, as one of the challenging vision tasks, aims to divide an image into certain coherent semantically meaningful parts. Fueled by deep learning, it has grown as the key enabler to cover navigation-related tasks in a unified way [7][8].
- An even higher potency of Convolutional Neural Networks (CNNs) arguably lies in the capacity to learn contexts and inter-relations. In our application domain, curbs appearing between roadways and sidewalks is one common property that is contextual information to be exploited despite the inherent variance in shapes, sizes and textures.
- Large-scale scene parsing datasets feature a high variability in capturing viewpoints (from road, sidewalks, and off-road areas) [9], which offer a broad range of images with assistance-related elements, supposing essential prerequisites to aid perception in VI individuals.

Based on these observations, we propose to seize pixel-wise semantic segmentation to support curb negotiation and traversability awareness as depicted in Fig. 1. This paper includes corresponding key contributions shaping our approach to this task, as well as novel results considerably extending previous preliminary works [4][6]:

- A real-world perception framework that unifies the

[1]Kailun Yang and Kaiwei Wang are with College of Optical Science and Engineering, Zhejiang University, Hangzhou, China {elnino, wangkaiwei}@zju.edu.cn;

[2]Luis M. Bergasa, Eduardo Romera and Rafael Barea are with Department of Electronics, University Alcalá, Madrid, Spain luism.bergasa@uah.es, eduardo.romera@edu.uah.es, rafael.barea@uah.es;

[3]Dongming Sun is with Department of Computing, Imperial College London, London, United Kingdom dongming.sun17@imperial.ac.uk.

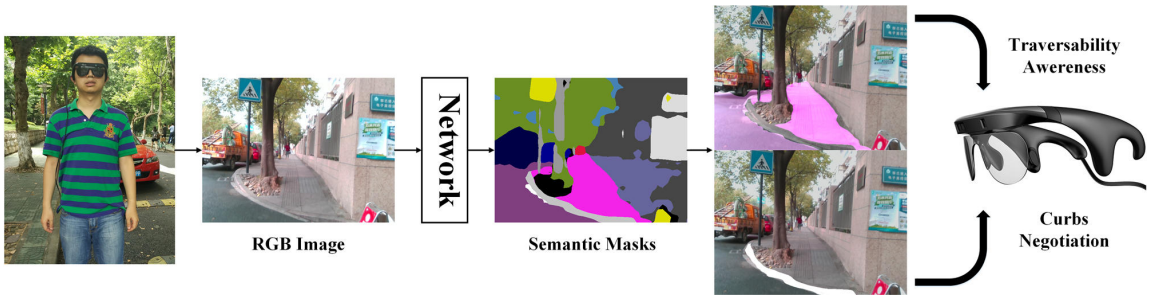


Fig. 2. Overview of the wearable navigation system.

hazardous curb detecting and traversable area parsing.

- A real-time semantic segmentation network to learn both global scene contexts and local textures without imposing any assumptions, while achieving better performance than traditional algorithms.
- A comprehensive set of experiments on two wearable navigation systems including the commercial smart glasses [10] and a highly customized prototype.

The remainder of this paper is structured as follows: Section II reviews related work that has addressed both traversability/curb detection and semantic segmentation for the visually impaired. In Section III, the framework is elaborated in terms of the perception system overview and the semantic segmentation architecture. In Section IV, the approach is evaluated and discussed as for real-time/real-world performance by comparing with traditional algorithms and state-of-art networks. Section V draws the conclusions and gives an outlook to future work.

II. RELATED WORK

Traversability detection was addressed by a vital part of proposals by adapting RANSAC algorithm to model the ground plane [5][6]. However, real-world ground areas are not always planar surfaces. Based on this knowledge, Stixel World [11] marked a significant milestone for flexibly representing traffic situations including the free road space as well as moving/static obstacles. On application side, possibilities were explored to leverage the Stixel-based techniques for self-driving cars, and adapt them into assistive technology for the VI. To overcome the limitation of incompatible assumptions across application domains, [3] exploited 3D indoor geometry to compute ground-to-image transformation, [2] clustered the normal vectors in the lower half of the field view, while [4] integrated IMU observations along with vision inputs in a straightforward manner.

Curb reconstruction is another research topic to complement the Stixel-based representation with regard to the insensitivity to low-height occurrence [12] as shown in Fig. 1. LiDAR point cloud was incorporated to enhance vision-based systems, aimed to achieve long-range curb detection, e.g., 20m claimed in [13]. With the similar purpose, multi-cue fusion [14] established boundary models of normal vector, height and color respectively. While LiDAR-based results are visually appealing, far-away information are less desired for navigation assistance than nearby hazard awareness. Towards this objective, the pioneering effort [15] demonstrated sufficient evidence for the presence of curbs and steps in simply one situation, forgetting to ensure the robustness across a broad spectrum of real-world scenarios. In past years, the proliferation of cost-effective depth sensors facilitated the evaluation of pavement unevenness and terrain roughness by computing surface normal vectors [16], which is arguably more suitable than cost-prohibitive LiDARs for wearable navigation systems. A more recent example could be [17], which consolidated a wearable technology to negotiate surface discontinuities using per-image classification. However, in complex urban areas, reliability of these systems are heavily influenced by the diverse street configurations, different materials/textures and illumination variations, let alone the viewpoint changes imposed by the wearable devices.

Semantic segmentation becomes visible and viable as an extremely powerful approach to provide a reliable generalization capacity with dense per-pixel predictions. However, the topic to leverage pixel-wise semantic segmentation to assist the visually impaired has not been practically studied. For prosthetic vision, a computer system [18] was reported to aid in obstacle avoidance through semantic labeling. Although related, the produced stimulation pattern can be thought of as a low resolution, low dynamic range, distorted image, which is insufficient for our task. Another piece of related work [19] has been recently presented to identify the most walkable direction for outdoor navigation. While inspiring, this work focused on the tracking of a safe-to-follow object by providing only sparse bounding-box semantic predictions, and hence cannot be directly used for upper-level reasoning tasks. Although sporadic efforts have been made along this line, these approaches are unable to run in real time, which is a critical issue for blind assistance. Additionally, they have not been thoroughly tested in the real world. Based on this notion, we attempt to customize real-time semantic segmentation unifying the perception of hazardous curbs beyond traversability, and offer an in-depth evaluation, focusing on a numerical analysis of real-world performance, followed by qualitative results as well as discussions.

III. APPROACH

A. Perception Framework Overview

To make the following explanations clear, we adopt the pair of smart glasses [10] worn by the user (see Fig. 2) as an instance. It is a commercialized product that aids obstacle avoidance during indoor/outdoor navigation, which is widely used in China by visually impaired pedestrians.

Following the trend of using head-worn glasses [6] to acquire environment information and interact with visually impaired pedestrians, we also design a customized prototype as shown in Fig. 3, which is comprised of a stereo camera attached with polarization filters [4]. In this work, the polarimetric information are not used so as to achieve a fair comparison with the smart glasses. The 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 pixel-wise semantic segmentation. As for the depth images, which are acquired with the combination of active speckle projecting and passive stereo matching, they enable a higher level of 3D pointcloud-based obstacle avoidance [6] that is robust yet insensitive to low-height hazardous curbs. Comparatively, for the customized prototype, the depth information are generated through purely large-scale stereo matching, designed originally for water hazard segmentation [4].

Fig. 3. The customized prototype.

For both navigation assistance systems, results regarding traversability awareness and curbs detection are determined by directly using the semantic segmentation output as the base for upper-level assistance, with which feedback are delivered through the bone conducting earphones. This is important as visually impaired pedestrians need to continue hearing environmental sounds when navigating different walkways and the bone conducting interface allows them to hear a layer of augmented acoustic reality that is superimposed on the environmental sounds, which are expected by the users to perceive the existence and direction of curbs, such that hazardous situations could be avoided or they could safely walk along the sidewalk.

B. Real-time semantic segmentation architecture

Up until very recently, the applicability of per-pixel semantic scene parsing is questioned due to speed. However, efficient semantic segmentation has been a heavily researched topic over the last two years, with the emergence of deep architectures [7][20][21] that could fulfill full segmentation in soft real time. These advances have empowered the utilization of semantic segmentation in time-critical applications like navigation assistance. In this research, to leverage the success in segmenting a broad spectrum of scenes and speeding up for semantic perception, the architecture is designed according to the SegNet-based encoder-decoder architectures like ENet [20] and our previous ERFNet [7]. In FCN-like architectures, featuremaps from different layers need to be fused to generate a fine-grained output. As expanded in Fig. 4, our approach in contrast uses a more sequential architecture based on an encoder producing down-sampling featuremaps, and a subsequent decoder that up-samples featuremaps to match input resolution. Table I also gives a detailed description of the integrated architecture. In general, the residual layer adopted in current networks has two instances: the bottleneck version and the non-bottleneck design. Based on 1D spatial factorizations of the convolutional kernels, "Non-bottleneck-1D" (non-bt-1D) was redesigned as an alternative residual layer in our previous work [7], successfully striking a rational balance between the efficiency of bottleneck and the learning capacity of non-bottleneck. Thereby, in order to extract featuremaps, an efficient use of minimized amount of residual layers is enabled with a maximized trade-off between inference speed and segmentation accuracy.

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	40×30
	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 featuremap	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 robust segmentation of street-level scene elements such as hazardous curbs and sidewalks, we attach a different decoder with respect to the previous work. This critical modification helps to collect more contextual information while minimizing the sacrifices of learning textures. Global context information is of cardinal significance to aid navigation at complex metropolitan areas. Detailedly, two common issues are worthwhile to highlight for context-critical blind assistance. Firstly, context relationship is universal, especially for street-level scene understanding. If the network mis-predicts curbs on crosswalks, the VI would be left vulnerable in the dynamic environments given such feedback. The prior knowledge should be learned by the data-driven approach that curbs are seldom over crosswalks. Secondly, when navigating the sidewalks or crossing the roads, the scene elements such as crosswalks, crossing lights, pedestrians, vehicles and the hazardous curbs will exhibit arbitrary sizes observed from the sensor perspective. Navigation assistance system should pay a lot attention to different sub-regions that contain inconspicuous-category stuff.

Learning more relationship between scene categories by

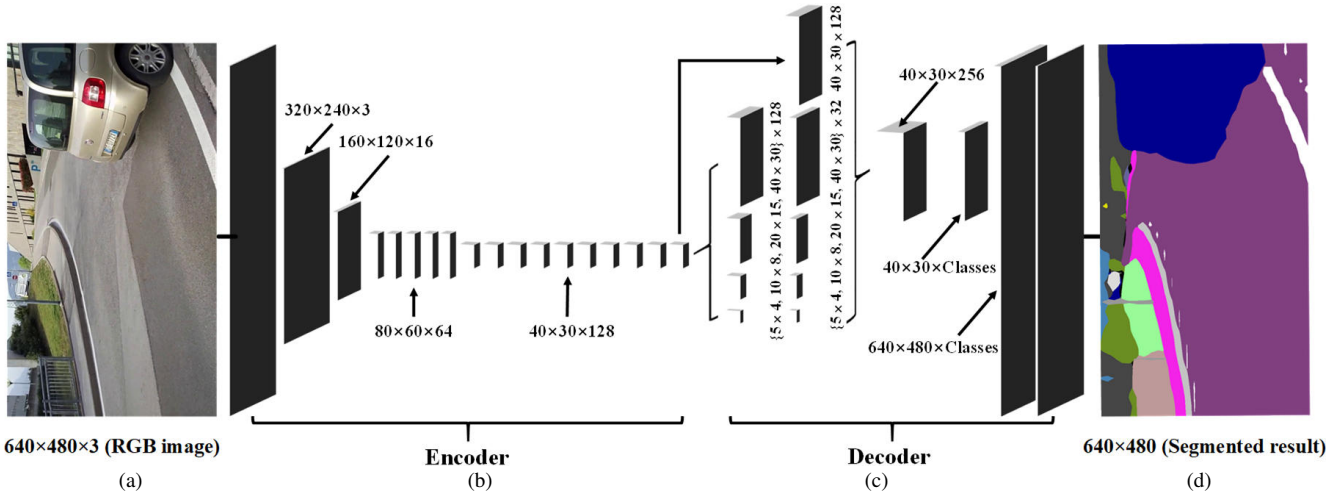


Fig. 4. The proposed architecture. From left to right: (a) Input, (b) Encoder, (c) Decoder, (d) Prediction.

exploiting more context is a promising approach to mitigate these risks. Bearing the goal of helping VI pedestrians in mind, the reconstruction of the decoder architecture follows the pyramidical pooling module as introduced by PSP-Net [22]. This module is leveraged to harvest different sub-region representations, followed by up-sampling and concatenation layers to form the final feature representations. In this manner, local and global context information are carried from the pooled representations at different locations. By fusing features under a group of different pyramid levels, the output of different levels in this pyramidical pooling module contains the featuremap from the encoder with varied sizes. With the aim of maintaining the weight of global feature, a convolution layer is appended after each pyramid level to point-wisely reduce the dimension of context representation to $1/N$ of the original one if the level size of the pyramid level is N. As for the situation in Fig. 4c, the level size N equals to 4 and we decrease the number of featuremaps from 128 to 32. Subsequently, the low-dimension featuremaps are directly up-sampled to retrieve the same-size features as the original featuremap through bilinear interpolation. Overall, Fig. 4 contains a depiction of the featuremaps generated by each of the block in our architecture, from the RGB input to the pixel-level class probabilities and final prediction.

IV. EXPERIMENTS

Experiments setup. The experiments are performed in public spaces around Westlake, the Yuquan Campus and the City College at Zhejiang University in Hangzhou, and the Polytechnic School at University of Alcalá in Madrid. We captured metropolitan scenes using two wearable navigation systems including the smart glasses commercially available at [10], and the customized prototype that was also previously used in [4] to detect water puddles by incorporating per-pixel polarimetric measurements. To facilitate fair comparison, polarization information are not used in this work. In this fashion, a real-world egocentric vision dataset is collected, which has 100 images across various scenarios with pixel-wise ground truth including 50 images captured by the glasses and 50 images captured by the prototype. This allows us not only to evaluate on the challenging large-scale Mapillary dataset [9], but also to analyze real-world performance using the curbs dataset that can be accessed at [24]. The metrics in this paper correspond to Intersection-over-Union (IoU) and Pixel-wise Accuracy (P-A) that are prevailing in semantic segmentation challenges.

Real-time performance. The total computation time for a single frame at the resolution depicted in Fig. 4/Table I is 13ms, mostly on semantic segmentation. In this sense, the computation cost is saved to maintain a reasonably qualified refresh-rate of 76.9FPS on a processor with a single cost-effective GPU GTX1050Ti. This inference time demonstrates that it is able to run our approach in real time, while allowing for additional time for acoustic feedback [4][6]. In addition, on an embedded GPU Tegra TX1 (Jetson TX1) that enables higher portability and consumption of less than 10 Watts at full load, our approach achieves approximately 22.0FPS.

Training setup. The challenging Mapillary Vistas dataset [9] is chosen as it consists of many traversability-related object classes including curbs, spanning a broad range of outdoor scenes on different roadways or sidewalks, which corresponds to the deployment scenario of both navigation assistance systems. Additionally, it attains vast geographic coverage, containing images from different continents. This is important to enhance reliability because curbs are not exactly the same in different streets and countries [13]. In total, we have 18000 images for training and 2000 images for validation with pixel-exact annotations. To provide awareness regarding the scenes that visually impaired people care the most during self-navigation, the training involves 27 classes, including the most frequent classes and some traversability-related classes. These 27 classes cover 96.3% of labeled pixels, which still allows to fulfill semantic scene parsing. To robustify the model against the varied types of images from real world, a set of data augmentations are performed including horizontally mirroring with a 50% chance, jointly use of random cropping and scaling to resize the cropped region into 320×240 input images. Particularly, random rotation by sampling distributions from the ranges [-20^o, 20^o], color jittering from the ranges [-0.2, 0.2] for hue, [0.8,

TABLE II

ACCURACY ANALYSIS.

Network	Traffic light	Car	Road	Sidewalk	Curb	Building	Person	Sky	Vegetation	Terrain	Crosswalk	Mean-11	Mean-27
ENet [20]	24.97%	71.16%	82.54%	57.20%	32.95%	75.97%	32.60%	96.39%	81.13%	52.85%	50.99%	59.89%	33.59%
LinkNet [21]	34.55%	74.41%	83.95%	58.22%	37.06%	78.16%	42.27%	**97.16%**	83.25%	54.88%	51.87%	63.25%	39.39%
ERF-PSPNet	**37.06%**	**75.92%**	**85.92%**	**65.14%**	**42.92%**	**80.52%**	**49.93%**	96.47%	**84.06%**	**60.09%**	**59.97%**	**67.09%**	**48.85%**

(a) On Mapillary dataset [9] using Intersection-over-Union (IoU).
"Mean-11": mean IoU value of 11 navigation-related classes, "Mean-27": mean IoU value of all 27 classes used for training.

Approach	Navigation System	All Pixels	With Depth	Within 2m	2-3m	3-5m	5-10m
3D-RANSAC-F	Smart Glasses	63.34%	67.16%	25.14%	95.67%	92.73%	54.75%
	Customized Prototype	80.88%	88.50%	91.13%	94.34%	92.70%	77.41%
	In Total	75.64%	82.68%	87.21%	94.55%	92.70%	70.21%
FreeSpaceParse	Smart Glasses	80.63%	81.11%	88.93%	86.98%	91.31%	78.38%
	Customized Prototype	86.52%	87.52%	69.11%	93.44%	92.35%	83.29%
	In Total	84.76%	85.77%	70.29%	92.45%	92.04%	81.73%
ENet	Smart Glasses	74.84%	75.14%	80.88%	64.42%	73.29%	78.00%
	Customized Prototype	93.26%	93.62%	93.43%	93.89%	94.74%	92.21%
	In Total	87.76%	88.58%	92.69%	89.36%	88.40%	87.70%
LinkNet	Smart Glasses	93.14%	92.76%	97.27%	95.20%	93.65%	92.63%
	Customized Prototype	93.28%	93.47%	93.62%	92.59%	93.40%	95.03%
	In Total	93.24%	93.28%	93.84%	92.99%	93.47%	94.27%
ERF-PSPNet	Smart Glasses	**96.86%**	**96.59%**	**99.58%**	**99.04%**	**98.43%**	**96.68%**
	Customized Prototype	**97.32%**	**97.55%**	**97.40%**	**97.73%**	**98.34%**	**96.34%**
	In Total	**97.18%**	**97.29%**	**97.53%**	**97.93%**	**98.36%**	**96.45%**

(b) On Real-world Curbs Dataset [24] in terms of traversable area detection using Pixel-wise Accuracy (P-A).
"With Depth": Only the pixels with valid depth information are evaluated.

Network	Navigation System	All Pixels	With Depth	Within 2m	2-3m	3-5m	5-10m
ENet	Smart Glasses	32.28%	33.59%	5.14%	7.83%	26.99%	20.92%
	Customized Prototype	53.22%	53.86%	44.53%	52.00%	50.57%	69.95%
	In Total	46.47%	47.56%	42.96%	50.53%	41.50%	45.87%
LinkNet	Smart Glasses	31.06%	32.98%	1.90%	11.88%	17.42%	47.39%
	Customized Prototype	49.37%	50.14%	59.02%	36.66%	56.74%	69.44%
	In Total	43.07%	44.80%	56.74%	35.83%	41.62%	58.61%
ERF-PSPNet	Smart Glasses	**79.11%**	**78.88%**	**66.90%**	**78.46%**	**82.26%**	**82.24%**
	Customized Prototype	**76.57%**	**77.66%**	**72.20%**	**73.38%**	**77.98%**	**89.16%**
	In Total	**77.38%**	**78.04%**	**71.99%**	**73.55%**	**79.63%**	**85.76%**

(c) On Real-world Curbs Dataset [24] in terms of curbs detection using Pixel-wise Accuracy (P-A).

1.2] for brightness, saturation and contrast are also applied. Our model is trained using Adam optimization, initiated with a batch size of 15 and a learning rate of 5×10^{-5} that decreases exponentially across epochs. Following the weight determining scheme in [20] and the pre-training setup in [7], the training of the full network reaches convergence when focal loss [23] is adopted as the criterion:

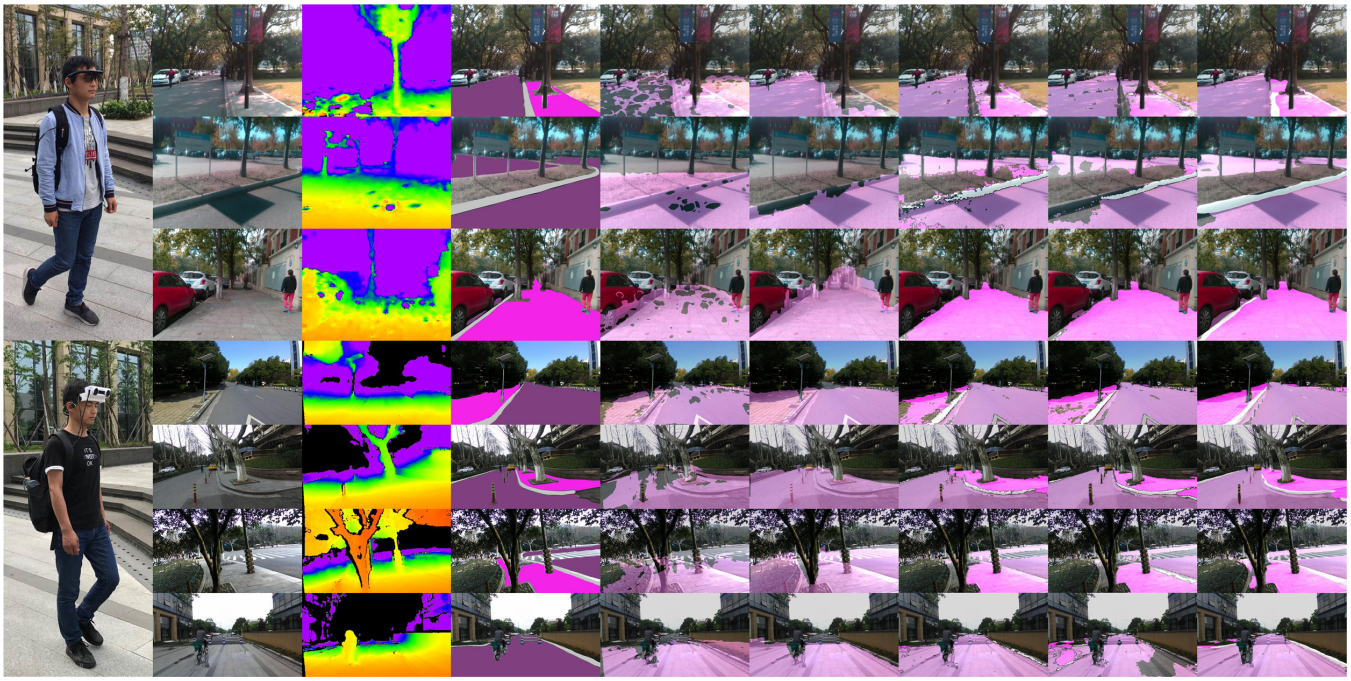
$$Focal_{loss} = \sum_{i=1}^{W}\sum_{j=1}^{H}\sum_{n=0}^{N}(1-\mathbf{P}_{(i,j,n)})^2\mathbf{L}_{(i,j,n)}log(\mathbf{P}_{(i,j,n)}) \quad (1)$$

where $\mathbf{P}$ is the predicted probability and $\mathbf{L}$ is the ground truth. The scaling factor $(1-\mathbf{P}_{(i,j,n)})^2$ suppressed heavily the loss contribution of correctly-segmented pixels (when $\mathbf{P}_{(i,j,n)}=0.9$, $(1-\mathbf{P}_{(i,j,n)})^2$=0.01). Comparatively, it suppressed lightly the loss contribution of wrongly-segmented pixels (when $\mathbf{P}_{(i,j,n)}=0.1$, $(1-\mathbf{P}_{(i,j,n)})^2$=0.81). In this way, the focal loss concentrates the training on wrongly-segmented pixels or hard pixels. We found this setting yields better results than conventional cross-entropy loss on Mapillary dataset, as it contains some less-frequent yet important classes such as traffic lights and hazardous curbs.

Segmentation accuracy. The accuracy of semantic segmentation is firstly evaluated on the challenging Mapillary dataset [9] by comparing the proposed ERF-PSPNet with CNNs in the state of the art including ENet [20] and LinkNet [21]. Table II(a) details the accuracy of 11 main navigation-related classes and the mean IoU values. It could be told that the accuracy of most classes obtained with the proposed ERF-PSPNet exceeds the existing architectures that are also designed for real-time applications. Our architecture has the ability to collect rich contextual information without major sacrifice of learning from textures. Accordingly, only the accuracy of sky is slightly lower than LinkNet, while most important classes for traversability awareness are both higher including road, sidewalk, curb, terrain and crosswalk. For other less frequent pathways, our approach also yields decent IoU value, e.g., bike lane (35.98%).

Real-world traversable area parsing. To analyze the major concern of detection performance for traversability awareness, we compare the traversable area parsing of our ERF-PSPNet to a traditional algorithm 3D-RANSAC-F [5], a Stixel-level segmentation pipeline FreeSpaceParse [3], as well as state-of-art architectures including ENet and LinkNet. Here, the traversable area involves road and sidewalk, excluding hazardous curbs. The pixel-wise accuracy on the real-world curbs dataset [24] over several ranges are collected: 0-2m, 2-3m, 3-5m and 5-10m, taking into account that short-range of ground area detection helps to determine the most walkable direction [4][19], while superior path planning could be supported by longer traversability awareness [6].

As manifested in Table II(b), 3D-RANSAC-F and FreeSpaceParse are based on depth segmentation by using the dense disparity map, which achieve decent accuracy between 2-5m because depth estimations within this range are



(a) Prototype (b) RGB (c) Depth (d) Annotation (e) 3D-R-F (f) FSP (g) ENet (h) LinkNet (i) Our approach

Fig. 5. Qualitative examples of the segmentation on real-world images produced by our approach compared with ground-truth annotation, 3D-RANSAC-F (3D-R-F) [5], FreeSpaceParse (FSP) [3], ENet [20] and LinkNet [21]. From left to right: (a) Wearable navigation system including the commercial smart glasses [10] and our customized prototype [4], (b) RGB image, (c) Depth image, (d) Annotation, (e) 3D-RANSAC-F (3D-R-F), (f) FreeSpaceParse (FSP), (g) ENet, (h) LinkNet, (i) Our approach.

quite dense with high degree of confidence. It is noteworthy that these two approaches both achieved higher accuracy with the customized prototype than the smart glasses. Admittedly, as depth information of the ground area captured with the smart glasses may be noisy and missing in dynamic environments, we implemented a RGB image guided filter [6] to fill holes before detection. However, the pair of smart glasses is designed to enable obstacle avoidance across indoor/outdoor environments through projected IR-texture stereo perception. While robust, its depth accuracy is less accurate in outdoor environments than the customized prototype using purely RGB stereo matching with longer baseline. This explains the accuracy gap between the glasses and the prototype.

As far as the deep learning-based approaches are concerned, they have the crucial advantages by exploiting a significant amount of data, thus eliminating the dependencies on assumptions. Since RGB-based segmentation is independent of depth information, there should be no significant difference between different navigation systems. Results collected with LinkNet and our ERF-PSPNet are as expected, and our approach outperforms them on both ranges, yielding high pixel-wise accuracy more than 97%. However, the accuracy gap still exists for ENet, because it suffers from limited learning capacity to generalize well in real world.

Real-world curbs detection. For the visually impaired, it is preferred to know that there are risks in some direction even if the per-pixel distinction is not exactly accurate. In this sense, the curbs segmentation results are still of great use even though they are less accurate than traversable area parsing. Interestingly, we observe a positive correlation between distance and accuracy obtained with our ERF-PSPNet in Table II(c). This is related to the Mapillary dataset used for training, in which most curbs are relatively farther regardless of viewpoints. However, such positive correlation does not hold for ENet and LinkNet, which reveals that LinkNet still suffers from limited capacity when learning/inferring less frequent scene classes. It is worthwhile to note the accuracy obtained with the smart glasses are slightly higher than that achieved with the customized prototype, because the aspect-ratio of images used for training is closer to aspect-ratio of images captured by the smart glasses than those from the prototype as displayed in Fig. 5, which may slightly bias the appearances of curbs to be analyzed, even though a group of data augmentations have already be performed. Still, our approach excels ENet and LinkNet, and in their cases, the accuracy obtained with the prototype are higher than the results of the glasses. This is mainly due to the low accuracy values at closer ranges (see Table II(c)), and the depth range of the smart glasses has been specially decreased to enhance obstacle avoidance as analyzed in [25]. It also implies that when the capacity is limited, less information are learned from the less-frequent classes than dominating traversable-area classes. In addition, our approach has the ability to gather diverse levels of context in the last layers, which also helps to learn from the close-range information that covers large portion of the image and requires the model to collect more contextual information for robust classification. In this regard, our approach is very suitable for navigation assistance because close-range hazard awareness is critical for the VI's safety, e.g., warnings of curbs within 3m given the height of head-worn devices. Fig. 5 exhibits the montage of pixel-wise results generated by our approach, LinkNet,

ENet, FreeSpaceParse and 3D-RANSAC-F. Qualitatively, our approach not only yields longer and more consistent segmentation which will definitely benefit the traversability awareness, but also retains the outstanding ability to perceive hazardous curbs within this framework.

V. CONCLUSIONS

Navigation assistance at metropolitan areas for the Visually Impaired (VI) is a necessary step to reach an optimal level of traffic flow, which will as well contribute to the improvements of transportation and vehicular safety. In support of that goal, we derive achievability results for perception of curbs and traversability by leveraging semantic segmentation, which has also played an important role in autonomous driving of Intelligent Vehicles (IV). The proposed approach has been evaluated on a large-scale challenging dataset and an ego-centric dataset, demonstrating the effectiveness in real-world assistance on two navigation systems.

We are aware of the plenty of room to further reinforce the robustness in unseen domains, and a field test with real VI users would garner more credibility. In the future, data augmentations and hierarchical structures would be analyzed in a systematic way on next generation of wearable prototypes that support higher portability, larger field of view and denser RGB-D perception, such that the framework would be ready for deployment in complex traffic situations. In addition, the navigational assistive framework would be incessantly enriched by chaining multi-sensor visual SLAM and life-long topological localization.

ACKNOWLEDGMENT

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 also 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. Mancini, E. Frontoni and P. Zingaretti, Mechatronic System to Help Visually Impaired Users During Walking and Running, IEEE Transactions on Intelligent Transportation Systems, 19(2), 2018, pp. 649-660.

[2] M. Martinez, A. Roitberg, D. Koester, R. Stiefelhagen and B. Schauerte, Using Technology Developed for Autonomous Cars to Help Navigate Blind People, In Computer Vision Workshop (ICCVW), 2017 IEEE International Conference on. IEEE, 2017, pp. 1424-1432.

[3] H. C. Wang, R. K. Katzschmann, S. Teng, B. Araki, L. Giarré and D. Rus, Enabling independent navigation for the 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.

[4] 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, 17(8), p. 1890, 2017.

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

[6] K. Yang, K. Wang, W. Hu and J. Bai, Expanding the detection of traversable area with RealSense for the visually impaired, Sensors 16(11), p. 1954, 2016.

[7] E. Romera, J. M. Alvarez, L. M. Bergasa and R. Arroyo, Erfnet: Efficient residual factorized convnet for real-time semantic segmentation, IEEE Transactions on Intelligent Transportation Systems, 19(1), pp. 263-272, 2018.

[8] K. Yang, K. Wang, L. M. Bergasa, E. Romera, W. Hu, D. Sun, J. Sun, R. Cheng, T. Chen and E. López, Unifying Terrain Awareness for the Visually Impaired through Real-Time Semantic Segmentation, Sensors, 18(5), p. 1506, 2018.

[9] G. Neuhold, T. Ollmann, S. R. Bulò and P. Kontschieder, The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes, In ICCV, 2017, pp. 5000-5009.

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

[11] H. Badino, U. Franke and D. Pfeiffer, The stixel world-a compact medium level representation of the 3d-world, In Joint Pattern Recognition Symposium, 2009, pp. 51-60.

[12] J. Siegemund, D. Pfeiffer, U. Franke and W. Förstner, Curb reconstruction using conditional random fields, In Intelligent Vehicles Symposium (IV), 2010 IEEE. IEEE, 2010, pp. 203-210.

[13] C. Fernandez, D. F. Llorca, C. Stiller and M. A. Sotelo, Curvature-based curb detection method in urban environments using stereo and laser, In Intelligent Vehicles Symposium (IV), 2015 IEEE. IEEE, 2015, pp. 579-584.

[14] L. Wang, T. Wu, Z. Xiao, L. Xiao, D. Zhao and J. Han, Multi-cue road boundary detection using stereo vision, In Vehicular Electronics and Safety (ICVES), 2016 IEEE International Conference on. IEEE, 2016, pp. 1-6.

[15] V. Pradeep, G. Medioni and J. Weiland, Piecewise planar modeling for step detection using stereo vision, In Workshop on computer vision applications for the visually impaired, 2008.

[16] K. Yang, K. Wang, R. Cheng and X. Zhu, A new approach of point cloud processing and scene segmentation for guiding the visually impaired, Iet International Conference on Biomedical Image and Signal Processing. IET, 2016, pp. 1-6.

[17] K. Y. Leong, S. Egerton and C. K. Chan, A wearable technology to negotiate surface discontinuities for the blind and low vision, In Life Sciences Conference (LSC), 2017 IEEE. IEEE, 2017, pp. 115-120.

[18] L. Horne, J. Alvarez, C. McCarthy, M. Salzmann and N. Barnes, Semantic labeling for prosthetic vision, Computer Vision and Image Understanding, 149, pp. 113-125, 2016.

[19] S. Mehta, H. Hajishirzi and L. Shapiro, Identifying Most Walkable Direction for Navigation in an Outdoor Environment, arXiv preprint arXiv:1711.08040, 2017.

[20] 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.

[21] A. Chaurasia and E. Culurciello, LinkNet: Exploiting encoder representations for efficient semantic segmentation, In Visual Communications and Image Processing (VCIP), 2017 IEEE. IEEE, 2017, pp. 1-4.

[22] 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.

[23] X. Y. Z. C. Riga, S. L. Lee and G. Z. Yang, Towards Automatic 3D Shape Instantiation for Deployed Stend Grafts: 2D Multiple-class and Class-imbalance Marker Segmentation with Equally-weighted Focal U-Net, arXiv preprint arXiv:1711.01506, 2017.

[24] Kaiwei Wang Team, Image Data Sets, http://wangkaiwei.org/projecteg.html, 2018.

[25] 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, 57(11), pp. 2809-2819, 2018.