Intersection perception through real-time semantic segmentation to assist navigation of visually impaired pedestrians

Kailun Yang[1], Ruiqi Cheng[1], Luis M. Bergasa[2], Eduardo Romera[2], Kaiwei Wang[1] and Ningbo Long[1]

***Abstract*— Intersection navigation comprises one of the major ingredient of Intelligent Transportation Systems (ITS) for Visually Impaired Pedestrians (VIP), who are the most vulnerable road users that should be protected with a high priority in metropolitan areas. Robotic vision-based assistive technologies sprung up over the past few years, which focused on specific scene objects using monocular detectors or depth sensors. These separate approaches achieved remarkable results with relatively low processing time, and enhanced the intersection perception to a large extent. However, running all detectors jointly incurs a long latency and becomes computationally prohibitive on wearable embedded systems. In this paper, we put forward to seize pixel-wise semantic segmentation to cover navigation-related perception needs in a unified way. This is not only critical to perceive crosswalk position (where to cross roads), traffic light signal (when to cross roads), but also to analyze the states of other pedestrians and vehicles (whether safe to cross roads). The core of our unification proposal is a deep architecture, aimed to attain efficient semantic understanding. A comprehensive set of experiments demonstrate the qualified accuracy over state-of-art algorithms while maintaining high inference speed on a real-world navigation assistance system.**

I. INTRODUCTION

Ambient smart living and Intelligent Transportation Systems (ITS) are becoming tightly intertwined [1] to enhance road safety assisted with robotic vision [2]. Intersections in complex metropolitan areas are one of the most hazardous where many accidents occur between turning-vehicles and pedestrians [3]. Rich functionalities have been included in mass-produced vehicles and transportation infrastructures [4], together with mobility aid for wheelchairs and individual travelers. In spite of the significant contributions of all these advances, there is still a long way to go towards the utopia of all traffic participants.

Arguably, most of the time ITS support able-bodied users to safely and efficiently use a transport system. Problems emerge when the user has some kind of disability, e.g., visual impairments. Precisely at urban intersections and roundabouts, Visually Impaired Pedestrians (VIP) encounter a diverse range of navigational challenges. There is a necessity to expand the coverage of assistance to help VIP crossing roads independently, which will also contribute to the improvements of transportation. Towards this end, a wide spectrum of tasks are concerned (see Fig. 1), with a vital part of vision-based proposals focused on crosswalk detection [3][5][6] and pedestrian crossing light detection [7][8]. In order to reduce traffic accidents during self-navigation, proof-of-concepts were also investigated to equip infrastructure-based pedestrian tracking [4] at signalized crosswalks, along with integration of wearable radar [9] to warn against collisions with vehicles, taking into consideration that fast-approaching objects are response-time critical.

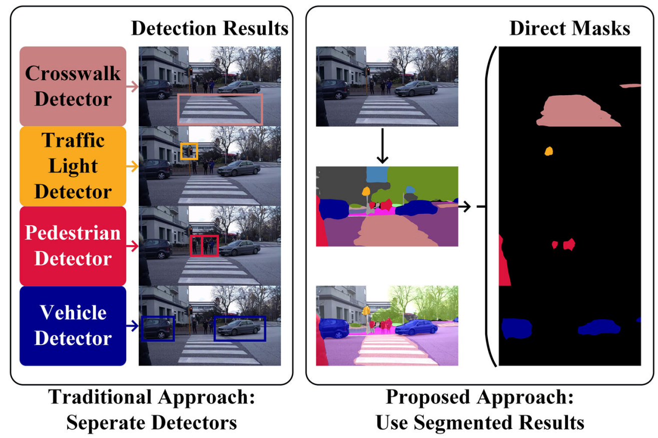


Fig. 1. Two approaches of perception in navigational assistance for visually impaired pedestrians at metropolitan intersections.

As a matter of fact, each one of these navigational tasks has been well resolved through its respective solutions. Despite the impressive strides towards higher mobility of VIP, a majority of processing pursues the sequential pipeline instead of a unified way, separately detecting different assistance-related scene elements. Thereby, it is computationally intensive to run different detectors together and the processing latency makes it infeasible within road crossing context. Illustratively, one of a pioneering work [7] recognizes traffic lights at about 5-10FPS, while delivering feedback in a few seconds. It sacrificed real-time performance by exploring temporal analysis for safety reasons. To locate crosswalks for transportation management system, [3] takes about 1.43s per frame based on MSER and ERANSAC. These approaches depend on further optimization to provide assistance at normal walking speed. A more recent example could be the wearable system reported in [6][8], which detects zebra

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.

[1]Kailun Yang, Ruiqi Cheng, Kaiwei Wang and Ningbo Long are with College of Optical Science and Engineering, Zhejiang University, Hangzhou, China {elnino, rickycheng, wangkaiwei, longningbo}@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.

crosswalks at about 15-30FPS, with additional 47ms to detect pedestrian crossing lights, let alone other processing components [10] that make it sub-optimal for real-time assistance on embedded platforms. In this sense, it is desirable to juggle multiple tasks simultaneously and coordinate all of the perception needs efficiently.

In order to close the gap, we derive insight from the field of autonomous driving, another safety-critical task that faces similar perception challenges, whose impressive developments could be leveraged for assistive intersection navigation given the following facts:

- Full pixel-wise semantic segmentation, as one of the challenging vision tasks, aims to partition an image into several coherent semantically meaningful parts. Fueled by deep learning, it has grown as the key enabler to cover navigation-related detection tasks in an end-to-end unified manner [11].
- An even higher potency of Convolutional Neural Networks (CNNs) arguably lies in the capacity to learn contexts and inter-relations. In our application domain, pedestrian crossing lights appearing above zebra crosswalks is one common property, which 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) [12], which offer a broad range of images with assistance-related intersection elements, supposing essential prerequisites to aid perception in visually impaired individuals.

Inspired by the synergy, we propose to seize pixel-wise semantic segmentation to provide a comprehensive set of assistive awareness, including crosswalk position (where to cross roads), traffic light signal (when to cross roads), as well as pedestrian and vehicle state (whether safe to cross roads). This paper considerably extends the previous work on traversability awareness [10] by including novel contributions and results that reside in the following aspects:

- A unification of intersection perception with regard to crosswalk detection, traffic light detection, pedestrian and vehicle detection.
- 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.
- A comprehensive set of experiments on a large-scale scene parsing dataset [12] and two real-world egocentric intersection datasets [6][8], by comparing with traditional algorithms and state-of-art networks.

The remainder of this paper is structured as follows. Section II reviews related work that has addressed both crosswalk detection, pedestrian traffic light detection for assistive navigation. In Section III, the proposed framework is elaborated in terms of the wearable navigation assistance system and the real-time semantic segmentation architecture. In Section IV, the approach is evaluated and discussed as for real-time/real-world performance by comparing to the most relevant approaches. Section V draws the conclusions and offers an outlook into what are expected in future work.

II. RELATED WORK

A large part of researches were dedicated to detecting merely one of landmarks at intersections, such as zebra crosswalks [3][5][6] or pedestrian crossing lights [7][8]. Comparatively, only a fraction of works have put efforts into the incorporation of crosswalk detection with crossing light detection. One of the earliest intersection assistance algorithm was proposed with analytic image processing [13]. It detects crossing lights in near-view images, where the light covers a dominant portion and no crosswalk exists, hence these two elements were not detected simultaneously. A robotic guide dog [14] was assembled with template matching-based crossing light detection and Hough transform-based crosswalk detection. However, this system was simply tested in one scenario, forgetting to guarantee the robustness across unseen situations. Another similar algorithm for intersection assistance based on RGB-D images [15] was specially designed to detect US crossing lights. In our application domain towards real-world assistance, the reliability should be ensured against the variety of street configurations, illumination changes, and even across continents. Pixel-wise semantic segmentation has come into view as an extremely powerful approach to provide a reliable generalization capability, as well as to detect multi classes of scenes simultaneously. However, the research topic to leverage semantic segmentation to assist VIP has not been investigated in complex traffic intersections/roundabouts. Our work aims to fill this gap.

III. APPROACH

A. Wearable assistive intersection navigation system

In this work, the main motivation is to design a prototype that should be wearable without hurting the self-esteem of VIP. With this target in mind, we follow the trend of using head-mounted glasses [10] to acquire environmental information and interact with VIP. As worn by the user at an urban intersection in Fig. 2, the pair of smart glasses is comprised of a RGB-D sensor of RealSense R200 and a set of bone conducting earphones.

This pair of smart 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, support a higher-level robust obstacle avoidance as previously presented in [10]. The crosswalk location, crossing light signal, and pedestrian/vehicle states are determined by directly using the semantic segmentation output as the base for assistive awareness, with which feedback are delivered through acoustic bone conduction.

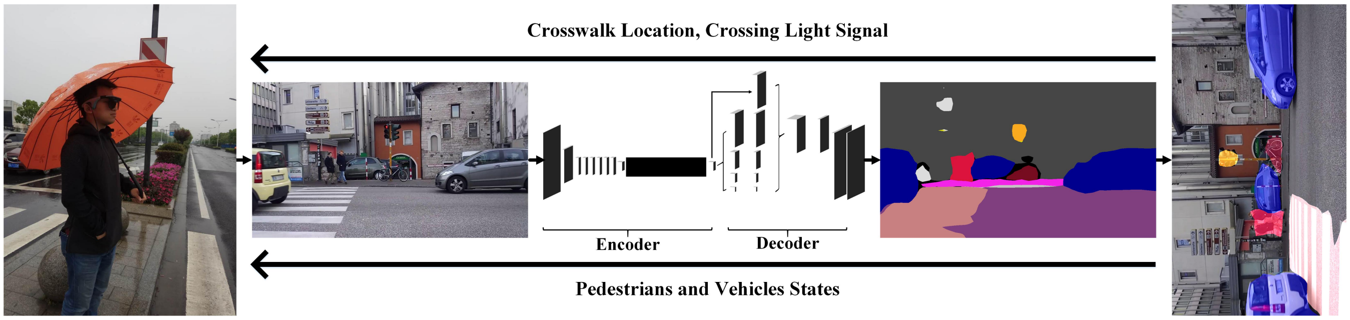


Fig. 2. Overview of the wearable navigation assistance system.

B. Real-time semantic segmentation architecture

Up until very recently, pixel-wise semantic segmentation was not deployable in terms of speed. However, a fraction of networks has focused on the efficiency by proposing architectures that could reach near real-time segmentation [11][16][17]. These advances have made possible the utilization of full scene segmentation in time-critical cases like blind assistance. 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 architectures like ENet [16] and our previous ERFNet [11]. In FCN-like architectures, feature maps from different layers need to be fused to generate a fine-grained output. 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. The integrated architecture can be also visualized in our work [2] with groups of collaborators whose goal is to create a unified terrain perception system. In this line, sequential/hierarchical architectures with spatially factorized convolution have been open-sourced at *https://github.com/elnino9ykl/ERF-PSPNet*.

For robust segmentation of intersection-centered scene elements, we attach a pyramidal pooling-based decoder [18], with the purpose to collect more contextual information while minimizing the sacrifices of learning textures. Global context information is of cardinal significance for navigational assistance at urban intersections. To detail this, two common issues are worthwhile to remark for context-critical blind assistance. First, context relationship is universal and important especially for complex scene understanding. If the network mis-predicts crosswalks on sidewalks, VIP would be left vulnerable in the dynamic environments. The common knowledge should be learned by the data-driven approach that crosswalks are seldom over sidewalks. Second, when crossing the roads, the scene elements such as crosswalks, crossing lights, pedestrians and vehicles are with arbitrary sizes from the sensor perspective. Navigation assistance system should pay much attention to different sub-regions that contain inconspicuous-category stuff. These risks could be mitigated by exploiting more context and learning more relationship between categories. Bearing the goal of helping VIP in mind, local and global context information are carried from the pyramidally harvested representations at different locations of the encoded featuremap with varied sizes. In summary, our ERF-PSPNet is built up through convolution factorization and pyramid representation allowing to learn high-level features hierarchically.

IV. EXPERIMENTS

Experiments setup. Datasets for evaluation include the challenging large-scale Mapillary dataset [12], and two real-world egocentric datasets [6][8] collected at urban intersections in Hangzhou, China and in Trento, Italy. A vital part of the images are captured by the wearable navigation assistance system, while others from smart phones share the same image style and quality. The metrics reported in this paper correspond to Intersection-over-Union (IoU) and Pixel-wise Accuracy (P-A) that are prevailing in semantic segmentation challenges, and two recall values in terms of stripe-level for crosswalk detection and instance-level for pedestrian crossing light detection.

Real-time performance. The total computation time of a single frame at the resolution of 320×240 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 GTX 1050Ti. This inference time demonstrates that it is able to run our approach in real time, while retaining additional time for acoustic feedback [10]. In addition, on an 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. When comparing the real-time performance with traditional detectors that focused on specific objects, our approach is the fastest as displayed in Table I, along with the forward passing time of state-of-art efficient architectures. At 320×240, our approach is slightly faster than ENet [16], even though LinkNet [17] is not able to be tested due to the inconsistent tensor sizes for down-sampling. At 640×480, our approach is also super fast. Still, our network achieves significantly higher accuracy than ENet and LinkNet, which will be detailed in following subsections.

Training setup. The challenging Mapillary Vistas dataset [12] is chosen as it consists of many navigation-related and intersection-centered object classes, spanning a broad range of outdoor scenes on different roadways or sidewalks, which corresponds to the usage scenario of the smart glasses. In addition, it attains vast geographic coverage, containing images from different continents. This is important to enhance reliability because zebra crosswalks and pedestrian crossing lights are not exactly the same in different countries.

TABLE I

REAL-TIME PERFORMANCE ANALYSIS.

Approach	Processing time
Crosswalk detection	
MSER and ERANSAC [3]	1.43s on Intel Core i7-3770
Bipolarity-based algorithm [5]	0.73s on Intel Core i7-3770
AECA algorithm [6]	33-67ms on Intel Atom x5-Z8500
Pedestrian crossing light detection	
Traffic light detection pipeline [7]	100-200ms on Nokia N95
PCL algorithm [8]	47ms on Intel Atom x5-Z8500
Semantic segmentation Networks are tested on a cost-effective GPU GTX1050Ti	
ENet [16]: 15ms at 320×240, 24ms at 640×480	
LinkNet [17]: Unable to be evaluated at 320×240, 32ms at 640×480	
Our ERF-PSPNet: 13ms at 320×240, 34ms at 640×480	

In total, we have 18000 images for training regardless of whether it contains crosswalks/crossing lights in the desired intersection/roundabout scenarios or not. Additionally, pixel-scale annotations of 2000 images are available for validation. Sharing the same spirit of past work [2] to unify awareness of the scenes that VIP care the most during self-navigation, we use 27 classes for training, including the most frequent classes and some assistance-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 group of data augmentations are performed including horizontally flipping with a 50% chance, jointly use of random cropping and scaling to resize the cropped regions into 320×240 input images. Random rotation by sampling distributions from the ranges $[-20^o, 20^o]$ is performed for wearable consideration. Color jittering from the ranges [-0.2, 0.2] for hue, [0.8, 1.2] for brightness, saturation and contrast are also applied to attain imaging invariance and robustness. Our model is based on the pre-trained encoder and training scheme in past work [2] while focal loss is used as the criterion.

Segmentation accuracy. The accuracy of semantic segmentation is firstly evaluated on the challenging Mapillary dataset [12] by comparing the proposed ERF-PSPNet with deep neural networks in the state of the art including ENet [16] and LinkNet [17]. Table II(a) details the accuracy of 11 frequent 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 intersection navigation are apparently higher including traffic light, car, person and crosswalk. For other less frequent vehicles/traffic participants, our approach also yields decent accuracy, e.g., truck (58.12%), bicycle (36.22%), motorcycle (39.79%), bus (61.35%) and rider (40.50%).

Real-world crosswalk detection. The crosswalk detection is evaluated on the Crosswalk Navigation dataset [6], which has 191 images with pixel-wise ground truth across 9 different scenarios for testing available at *http://wangkaiwei.org/projecteg.html*. This allows us to compare our approach with traditional approaches including the bipolarity-based algorithm [5], Adaptive Extraction and Consistency Analysis (AECA) algorithm [6], as well as state-of-art networks including ENet and LinkNet. Considering the sharp contrast in the boundaries of black-white stripes, [5] detected crosswalks by analyzing bipolarity of gray-scale histogram. However, the performance of the algorithm is sensitive to the pre-determined segmenting size of patches. Therefore, the crosswalks at far distances fail to be detected (see Fig. 3d), resulting a low accuracy and stripe-level recall as observed in Table II(b). Comparatively, AECA only extracts bright stripes of zebra crosswalks, thus its pixel-wise accuracy and IoU are unable to compare fairly with other approaches. It claimed to surpass bipolarity-based algorithm in terms of frame-level precision and recall. However, it is noticeable that not all of crosswalk stripes are included in detection results as displayed in Fig. 3e. Due to the incomplete detection, the close crosswalk stripes whose features are less consistent with most stripes may miss, which results in delivering confusing feedback as pointed out in [6].

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. Intriguingly, although LinkNet exceeds ENet on Mapillary dataset, only the recall is higher than ENet on the real-world dataset. ENet applied multiple dilated convolution by taking a wider context into account, while LinkNet only performed fixed ones. Accordingly, ENet outperforms LinkNet in terms of IoU, because close-range stripes' sizes vary greatly when crossing the roads, which requires the model to learn rich contextual information and these stripes cover most pixels. However, LinkNet has larger capacity and it surpasses ENet in terms of recall, which are largely contributed by relatively farther stripes. Still, our ERF-PSPNet excels on both metrics, although in some scenarios the pixel-wise accuracy are slightly lower than ENet/LinkNet because they sometimes tend to over-segment crosswalks, e.g., classify general road markings as zebra crosswalks, leading to inferior real-world performance. Fig. 3 exhibits the montage of detection results generated by our approach, bipolarity-based algorithm and AECA approach. Qualitatively, our approach yields longer and more consistent segmentation which will definitely benefit the assistive awareness at urban intersections.

Real-world pedestrian crossing light detection. For another critical task, pedestrian crossing light detection is evaluated on the real-world dataset [8]. This dataset contains several video clips captured in China (4867 images) and Italy (12913 images). A real-time PCL algorithm [8] detects lights based on HOG and SVM. It only segments bounding-box pedestrian region of the lights, relying on the HOG descriptor to classify candidates. In contrast, our approach detects not only pedestrian crossing lights but also other kinds of traffic lights, which arguably supports more comprehensive upper-level analysis and assistance. In order to facilitate fair comparison, we collected the instance-level recall as itemized in Table III, which is a very important parameter for time-critical blind assistance, relaxing the requirements

TABLE II

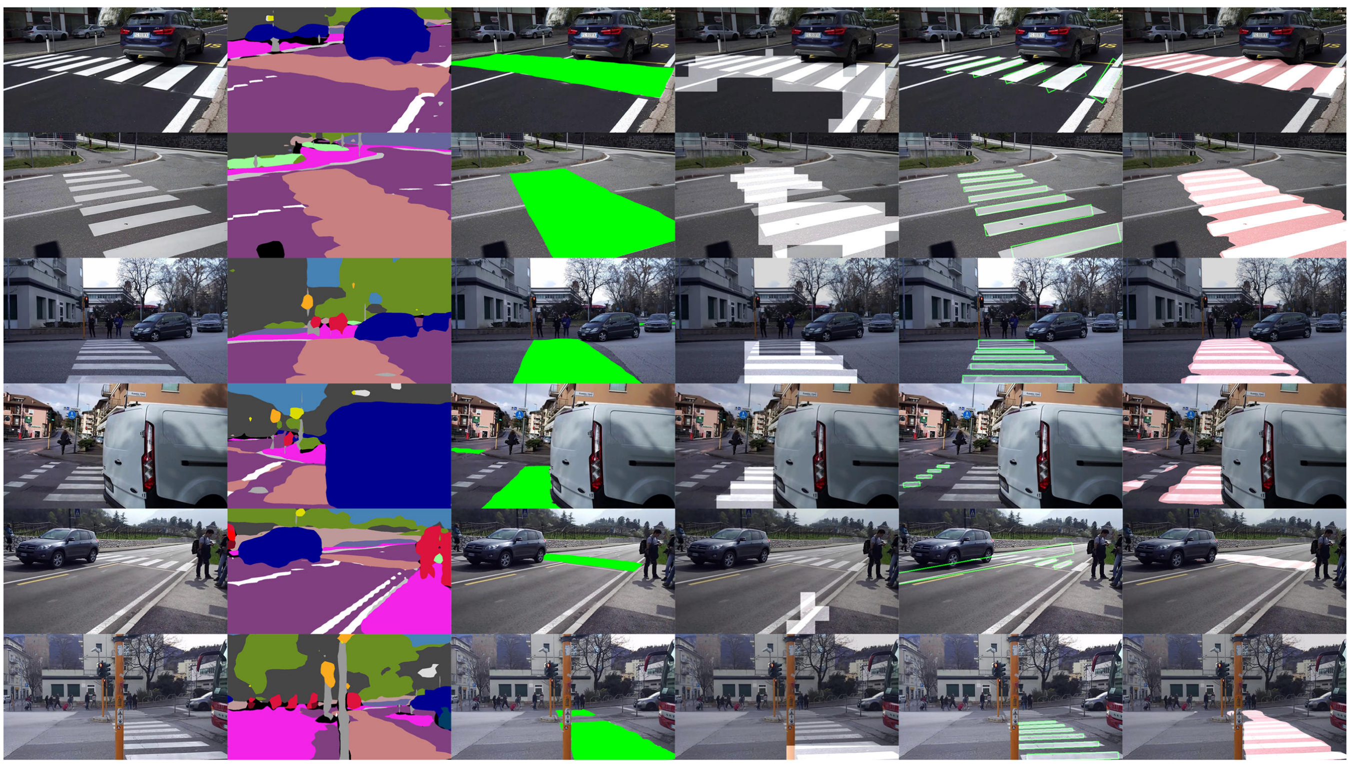
ACCURACY ANALYSIS.

Network	Traffic light	Car	Road	Sidewalk	Curb	Building	Person	Sky	Vegetation	Terrain	Crosswalk	Mean-11	Mean-27
ENet [16]	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 [17]	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 [12] 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.

Scenario	Bipolarity-based [5]			AECA [6]	ENet [16]			LinkNet [17]			Our approach		
	IoU	P-A	Recall	Recall	IoU	P-A	Recall	IoU	P-A	Recall	IoU	P-A	Recall
Scenario 1	64.48%	67.99%	45.00%	36.52%	87.24%	94.76%	75.00%	74.83%	**96.59%**	78.04%	**88.87%**	95.82%	**91.52%**
Scenario 2	33.05%	34.37%	16.78%	33.56%	75.70%	86.36%	69.13%	71.57%	89.80%	78.52%	**81.14%**	**94.02%**	**85.23%**
Scenario 3	15.83%	17.73%	17.19%	33.26%	69.87%	85.11%	70.31%	54.63%	86.11%	72.54%	**80.15%**	**90.39%**	**87.72%**
Scenario 4	9.16%	9.44%	9.09%	55.84%	66.07%	**94.07%**	**100.0%**	65.24%	86.78%	98.70%	**77.62%**	93.25%	**100.0%**
Scenario 5	0.00%	0.00%	0.00%	67.74%	42.05%	42.50%	48.39%	55.58%	**75.82%**	77.42%	**70.60%**	73.56%	**90.32%**
Scenario 6	52.94%	69.37%	63.64%	50.00%	57.01%	58.19%	69.09%	35.25%	52.53%	72.73%	**81.52%**	**85.43%**	**98.18%**
Scenario 7	25.96%	26.95%	27.34%	57.55%	72.14%	76.75%	66.91%	69.92%	**87.59%**	84.89%	**79.90%**	84.05%	**92.09%**
Scenario 8	0.00%	0.00%	0.00%	29.41%	88.97%	96.64%	64.71%	87.34%	96.67%	**88.24%**	**89.16%**	**97.97%**	**88.24%**
Scenario 9	73.92%	83.30%	95.63%	58.52%	64.64%	**98.35%**	98.25%	67.04%	93.93%	94.32%	**81.02%**	96.59%	**99.56%**
In total	50.38%	55.87%	38.73%	42.47%	70.86%	88.70%	75.90%	64.08%	88.63%	80.12%	**82.50%**	**92.83%**	**91.87%**

(b) On real-world Crosswalk Navigation dataset [6]. "P-A": Pixel-wise Accuracy.



(a) RGB image (b) Segmented masks (c) Annotation (d) Bipolarity-based (e) AECA algorithm (f) Our approach

Fig. 3. Qualitative examples of the zebra crosswalk detection on real-world images produced by our approach compared with ground-truth annotation, bipolarity-based approach [5] and AECA algorithm [6]. From left to right: (a) RGB image, (b) Segmented masks of ERF-PSPNet, (c) Annotation, (d) Bipolarity-based, (e) AECA algorithm, (f) Our approach.

TABLE III

INSTANCE-LEVEL RECALL

ON REAL-WORLD PEDESTRIAN CROSSING LIGHTS DATASET [8].

Approach	China dataset	Italy dataset	In total
PCL algorithm [8]	46.77%	64.71%	59.53%
ENet [16]	51.61%	83.67%	74.42%
LinkNet [17]	64.52%	93.84%	82.33%
Our approach	**75.81%**	**96.08%**	**89.77%**

of temporal analysis that hinders real-time feedback. We counted the pedestrian traffic lights for images at an interval of 100 frames of the datasets, having 62 lights in 48 frames of the China dataset and 153 lights in 129 frames of the Italy dataset. Numerically, the recall of our approach is the highest among these real-time algorithms. As far as the color signal is concerned, our approach achieves decent precision of more than 90% for red lights and more than 95% for green lights by setting thresholds in HSV space, given that the red and green PCL gather around specific values of Hue and Value [8]. To further improve the precision in future time, we aim to implement illumination-invariant image pre-transformation, as well as to incorporate near-infrared spectral information. It is also worthwhile to note

(a) RGB image (b) Segmented masks (c) PCL algorithm (d) ENet (e) LinkNet (f) Our approach

Fig. 4. Qualitative examples of the pedestrian crossing lights detection on real-world image produced by our approach compared with ground-truth annotation, PCL algorithm [8], ENet [16] and LinkNet [17]. From left to right: (a) RGB image, (b) Segmented masks of ERF-PSPNet, (c) PCL algorithm, (d) ENet (e) LinkNet, (f) Our approach.

that the recall values in Italy dataset are all higher than the results of China dataset. First, intersections in China dataset are more crowded and complex as shown in Fig. 4, which are inherently more difficult than images in Italy dataset. Second, in spite of being with a global reach, the Mapillary dataset for training contains more images from Europe than from Asia, which may slightly bias the appearances of objects to be analyzed. This explains the recall gap between two countries, even though our approach is already able to generalize far beyond its training data, manifesting qualified detection results across diverse scenarios.

V. CONCLUSIONS

Navigational assistance at urban intersections for Visually Impaired Pedestrians (VIP) is a necessary step to reach an optimal level of traffic safety, which is one major focus of Intelligent Transportation Systems (ITS). In this paper, we derive achievability results for unifying intersection-centered perception tasks by utilizing real-time semantic segmentation, which is able to render a comprehensive set of assistive awareness without incurring a long latency. The proposed approach has been evaluated on a large-scale challenging dataset and two egocentric datasets across different countries, demonstrating the effectiveness in real-world assistance on the wearable navigation system. Future works will involve panoramic image semantic segmentation and multi-modal sensory perception to constantly enhance the navigation assistive framework.

REFERENCES

[1] M.A. Taie, K.M. NasrEldin and M. ElHelw, “ITS navigation and live timetables for the blind based on RFID robotic localization algorithms and ZigBee broadcasting,” In *ROBIO*, 2013.

[2] 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*, 2018.

[3] Y. Zhai, G. Cui, Q. Gu and L. Kong, “Crosswalk detection based on MSER and ERANSAC,” In *ITSC*, 2015.

[4] D.F. Llorca, R. Quintero, I. Parra, R. Izquierdo, C. Fernandez and M.A. Sotelo, “Assistive pedestrian crossings by means of stereo localization and rfid anonymous disability identification,” In *ITSC*, 2015.

[5] M.S. Uddin and T. Shioyama, “Bipolarity and projective invariant-based zebra-crossing detection for the visually impaired,” In *CVPRW*, 2015.

[6] R. Cheng, K. Wang, K. Yang, N. Long, W. Hu, H. Chen, J. Bai and D. Liu, “Crosswalk navigation for people with visual impairments on a wearable device,” *J. Electron. Imaging*, 2017.

[7] J. Roters, X. Jiang and K. Rothaus, “Recognition of traffic lights in live video streams on mobile devices,” *IEEE Trans. Circuits Syst. Video Technol.*, 2011.

[8] R. Cheng, K. Wang, K. Yang, N. Long, J. Bai and D. Liu, “Real-time pedestrian crossing lights detection algorithm for the visually impaired,” *Multimed. Tools Appl.*, 2017.

[9] N. Long, K. Wang, R. Cheng, K. Yang and J. Bai, “Fusion of millimeter wave radar and RGB-depth sensors for assisted navigation of the visually impaired,” In *Proc. of SPIE*, 2018.

[10] K. Yang, K. Wang, W. Hu and J. Bai, “Expanding the detection of traversable area with RealSense for the visually impaired,” *Sensors*, 2016.

[11] E. Romera, J.M. Alvarez, L.M. Bergasa and R. Arroyo, “Erfnet: Efficient residual factorized convnet for real-time semantic segmentation,” *IEEE Trans. Intell. Transp. Syst.*, 2018.

[12] G. Neuhold, T. Ollmann, S.R. Bulò and P. Kontschieder, “The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes,” In *ICCV*, 2017.

[13] T. Shioyama, H. Wu, N. Nakamura and S. Kitawaki, “Measurement of the length of pedestrian crossings and detection of traffic lights from image data,” *Meas. Sci. Technol.*, 2002.

[14] Y. Wei, X. Kou and M.C. Lee, “A new vision and navigation research for a guide-dog robot system in urban system,” In *AIM*, 2014.

[15] S. Wang, H. Pan, C. Zhang and Y. Tian, “RGB-D image-based detection of stairs, pedestrian crosswalks and traffic signs,” *J. Vis. Commun. Image R.*, 2014.

[16] A. Paszke, A. Chaurasia, S. Kim and E. Culurciello, “Enet: A deep neural network architecture for real-time semantic segmentation,” *arXiv*, 2016.

[17] A. Chaurasia and E. Culurciello, “LinkNet: Exploiting encoder representations for efficient segmentation,” In *VCIP*, 2017.

[18] H. Zhao, J. Shi, X. Qi, X. Wang and J. Jia, “Pyramid scene parsing network,” In *CVPR*, 2017.