Robustifying Semantic Cognition of Traversability across Wearable RGB-Depth Cameras

KAILUN YANG[1], LUIS M. BERGASA[2], EDUARDO ROMERA[2], AND KAIWEI WANG[1,*]

[1] State Key Laboratory of Modern Optical Instrumentation, Zhejiang University, Hangzhou 310027, China
[2] Department of Electronics, University of Alcalá, Madrid 28805, Spain
[*] Corresponding author: wangkaiwei@zju.edu.cn

Compiled March 21, 2019

Semantic segmentation represents a promising means to unify different detection tasks, especially pixel-wise traversability perception as the fundamental enabler in robotic vision systems aiding upper-level navigational applications. However, major research efforts are being put into earning marginal accuracy increments on semantic segmentation benchmarks, without assuring the robustness of real-time segmenters to be deployed in assistive cognition systems for the visually impaired. In this paper, we explore in a comparative study across four perception systems, including a pair of commercial smart glasses, a customized wearable prototype and two portable RGB-Depth (RGB-D) cameras that are being integrated in the next generation of navigation assistance devices. More concretely, we analyze the gap between the concepts of "accuracy" and "robustness" on the critical traversability-related semantic scene understanding. A cluster of efficient deep architectures is proposed, which are built using spatial factorizations, hierarchical dilations and pyramidal representations. Based on these architectures, this research demonstrates the augmented robustness of semantically traversable area parsing against the variations of environmental conditions in diverse RGB-D observations, and sensorial factors such as illumination, imaging quality, field of view and detectable depth range.

© 2019 Optical Society of America

http://dx.doi.org/10.1364/ao.XX.XXXXXX

1. INTRODUCTION

The increasing demand for traffic safety has spurred efforts from both academia and industry to jointly develop technologies for robotics, autonomous vehicles and vulnerable road users. In this line, one of the concrete examples of this alliance is the resurgence of young companies collaborating with universities to build robotic society where humans and robots interact seamlessly, like KR-VISION [1] that cares about the needs of visually impaired individuals by prototyping generations of assisted cognition and navigation systems.

Towards this end, roboticists and entrepreneurs are endeavoring to apply computer vision and 3D imaging techniques, which open the new avenues to solve numerous problems and surpass human-level performance, especially in visual perception with the advent of powerful object detectors [2] and semantic segmenters [3]. Precisely, the segmentation process, posed as per-pixel prediction to divide observed scenes into semantic regions, has been promising for unifying different detection tasks that are needed for safe navigation [4, 5].

Nowadays, Convolutional Neural Networks (CNNs) stand out over state-of-the-art solutions for Semantic Segmentation (SS). The resounding success is not only due to the continuous accumulation and expansion of per-pixel annotated datasets, but also the affordable computational resources and mobile platforms like Nvidia Jetson TX1/TX2. In particular, the large datasets such as Cityscapes [6], Mapillary VISTAS [7], PASS [8] and RANUS [9], not only feature a high variability in capturing viewpoints (from road, sidewalks, and off-road), but also span a broad range of scenes on different pathways, which extremely facilitates training of deep models for pixel-wise SS in the corresponding context of assisted navigation. Additionally, those mobile platforms are appealing for real-world applications as they consume less power and are more compact compared to desktop workstations counterparts.

As a matter of fact, the choice of CNN architecture does play a crucial role for SS. However, most of the efforts are spent in pursuing finer quality and marginal accuracy boosts with sophisticated modules [10–13], forgetting that SS algorithms must be

efficient and deployable in diverse navigation assistance systems (see Figure 1), and face real-world scenarios observed through different cameras. The question then naturally arises, is the SS approach robust enough? From a robotic perspective, this is often related to the safety of vision-based wearable cognition systems, since the SS process will serve as the key enabler to detect traversability, based on which the user with the semantic foresight can find the walkable directions and navigate through obstacles independently [14]. On the other hand, reliable segmentation in unseen domain is challenging, meanwhile attaining the robustness across diverse observations of wearable cognition systems will be highly difficult.

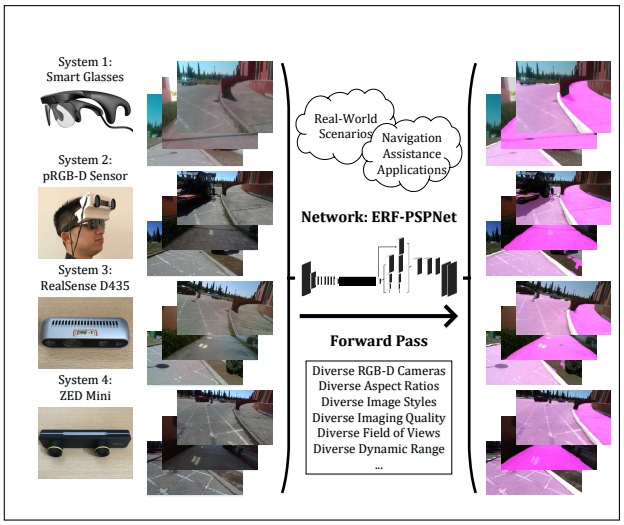

Fig. 1. Overview of the study on the robustness of semantic traversability perception across navigation assistance systems.

Based on these observations, we aim to address the dilemma, by exploring in a comparative study across four navigational perception systems, including a pair of popularly used smart glasses, a customized prototype and two miniaturized RGB-Depth (RGB-D) sensors that are being integrated in the next generation of wearable devices. More concretely, this work not only introduces new concepts of factorized convolutions, hierarchical dilations and pyramid representations that constitute the technical gists of our deep architectures to fulfill real-time end-to-end predictions, it also highlights the essential role of data augmentation in robustifying the perception systems. A comprehensive set of experiments demonstrates that with appropriate data augmentation techniques regarding geometry (position and shape) and texture (color and illumination), the models based on our architecture can be robust to the changes of multiple environmental and sensorial factors.

The paper proceeds as follows: Section 2 reviews the literature mainly related to traversability perception and SS taking into account the efficiency and robustness. In Section 3, the framework is elaborated in terms of system descriptions, architecture designs and data augmentations. In Section 4, we present the comprehensive experiments and discussions. Finally, Section 5 concludes the paper with scope of future directions.

2. RELATED WORK

Traversability detection is classically denoted as the segmentation of traversable areas. A vital part of early attempts modeled the ground plane by using growing-based algorithms, or adapting RANdom Sampling Consensus (RANSAC) [15, 16]. However, real-world ground areas are not always planar surfaces or ideally flat. Inspired by this notion, Stixel World [17] marked a significant milestone in the context of intelligent vehicles. The original motivation was to detect non-flat road surfaces for autonomous or assisted driving, which allows for flexible and efficient representation of traffic situations including free space and moving obstacles.

On application side, adaptation strategies [18–20] were explored to leverage the Stixel-based technologies for self-driving cars, and integrate them in wearable navigation assistance systems. The spirit behind is that the representation can be explored to generate acoustic feedback by sonifying stixels [18] or haptic feedback [19] with vibrating motors, such that visually impaired users can perceive the surroundings without excess burdens. To overcome the limitations of incompatible assumptions across application domains, [18] clustered the normal vectors in the lower half of the field of view, [19] measured the 3D geometry of the surrounding layout, while [20] combined additional IMU observations with vision inputs in a straightforward fashion. However, Stixel-based representation is not sensitive to low-height objects, which means semantic foresight of low-lying obstacles can not be easily gained [21]. Moreover, the concept of traversability perception should not be confined to the segmentation of traversable areas and obstacles. Meanwhile, the uneven parts of roadways can also be taken into consideration, such as semantic awareness of hazardous curbs, which is ardently desired by visually impaired pedestrians as underlined by our previous field test [5].

Semantic segmentation becomes visible as a promising approach to provide a potential generalization and viable unification capacity. Intuitively, the detection of roadways, sidewalks, various terrains and intersections/roundabouts can be covered in a unified SS framework [5, 22–24]. Although the detection of curbs has not been practically embodied in any existing SS framework, we argue that it can also be solved by a semantic segmenter. Thereby, in this paper, the performance of pixel-level curb detection is studied together with traditional traversable area parsing.

Researchers have already put computer vision-oriented robots into perspective by formulating terrain traversability analysis as a pixel-wise SS problem. According to the target environment of a search and rescue robot [25], a Fully Convolutional Network (FCN) [3] was fine-tuned, such that the SS model is geared towards the cross-season off-road setting. In order to close the gap between semi-autonomous and fully autonomous driving capabilities, a FCN was trained to gain semantic understanding with a GoogLeNet architecture [26], which was then fused with the 3D Stixel-based representation to detect small-sized and unexpected road hazards. Their approach provides a high classification accuracy at a relative low computational cost and GPU memory demands. Such a reasonable trade-off for self-driving cars has also been achieved by examining the possibility of building smaller architectures, to name a few efficient networks, SegNet [27], ENet [28], LinkNet [29], ERFNet [30], ICNet [31], SQNet [32], DSNet [33], ContextNet [34], FastSCNN [35], and VeryFastNetwork [36]. In this paper, we propose a cluster of SS architectures that strike excellent trade-off convincingly appealing for real-world applications, with special insights into wearable navigation assistance.

While unsupervised or weakly-supervised segmentation frameworks [37] generally entail the use of multiple sensors and suffer from the poor universality of models, currently most

segmentation networks must learn from labeled data in a supervised way to achieve top accuracy. Datasets like Cityscapes [6] and Mapillary VISTAS [7] have thousands of images, but even their diversity does not guarantee top performance of contemporary segmenters when ported on an unknown system against unseen real-world domains. Under the vital topic of visual domain adaptation, alternative approaches have been motivated to facilitate adversarial learning in feature or output space [38], such that cross-domain data could be jointly used. However, such approaches are hard to train as they rely on relatively unstable Generative Adversarial Network (GAN) [39] setups. Additionally, under common cases, we have no access to the deployment environment of the individual consumer-oriented systems. Despite these recent advances, it is still unclear how SS networks generalize in an unseen domain across wearable systems.

Robustness study of SS is of immense importance, but it represents a challenging and so far not fully investigated topic. Whereas CNN-based semantic perception can be applied to the fields of assisted navigation as conceptually validated in some wearable/mobile systems [5, 14, 40], what dominate in practice are still depth-based segmentation pipelines [15, 16, 18–20]. To address this dilemma, a large body of work used an additional sensor such as LiDAR towards robust semantic scene understanding [41]. Another cluster of frameworks incorporated a multi-stream deep neural network to learn features from complementary modalities and spectra, each of which is specialized in a subset of the input space, such as depth cues in [42], near-infrared images in [9], and predicted per-pixel polarization information in [43]. Although these works did confirm the benefits of using multi-modal data for deep scene parsing, non-RGB spectrum data are not always available in robotics applications.

In order to overcome the impact of non-idea weather conditions and isolate the perception of scenes from environmental effects, N. Alshammari *et al.* [44] used an illumination invariant transformation, and notably improved SegNet [27]-based segmentation performance. Similarly extending semantic scene understanding to adverse conditions with degraded visibility, C. Sakaridis *et al.* [45, 46] made attempts to gradually adapt SS models from synthetic domains to foggy and nighttime driving scenes, clearly evidencing the generalization benefits.

On the other hand, the robustness of classical SS architectures and commonly-used additional components have been rigorously evaluated, especially to adversarial examples [47]. As far as range-related performance is concerned, G. L. Oliveira *et al.* [48] explicitly collected a dataset to measure the robustness of their human body part segmentation networks when exposed to multiple scales. While inspiring, the orientation of human body in their test is invariant, which is apparently less complex than real-world traversable area parsing. O. Zendel *et al.* [49] assessed the real-world applicability of semantic segmentation algorithms against challenging data, especially those with potential hazards by embracing the global diversity of traffic situations extracted from dashcam video materials. M. Larsson *et al.* [50] created cross-season correspondence datasets to facilitate further research on making SS CNNs more robust to seasonal and weather changes.

In summary, even though this research area is becoming active, there are still lots of gaps. Specifically, the robustness across diverse RGB-D observations from wearable semantic cognition systems has not been thoroughly studied. Our work comes to fill this gap.

3. APPROACH

A. Perception Framework Overview

The semantic perception framework of traversable areas and hazardous curbs has been integrated into the KR-Vision Smart Glasses [1] as an instance depicted in Figure 2. It is a commercialized product that aids obstacle avoidance during indoor/outdoor navigation based on RealSense R200 [52]. We also design a customized prototype (pRGB-D Sensor [20]), which is comprised of a ZED stereo camera [51] attached with polarization filters. In addition, we have two portable RGB-D cameras including RealSense D435 [53] and ZED Mini, which is part of the wearable mixed-reality system [54] bringing the best of virtual and augmented reality together.

To make the following explanations clear, we adopt the pair of Smart Glasses [1] worn by the user (see Figure 2) as an instance to describe the role of a RGB-D camera-based perception framework in a navigation assistance system. 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 pixel-wise SS. As for the depth images, they enable a higher-level of pointcloud-based obstacle avoidance [16]. In this RGB-D perception pipeline, the depth images can also be used to detect ground areas by using RANSAC-based algorithm [15] or Stixel-based representation [19], which will also be studied by ensuring fair comparability.

B. RGB-D Sensor Systems

The Smart Glasses (RealSense R200 integrated) and RealSense D435 utilize active stereo, an extension of the traditional passive stereo approach in which a pattern is projectively texturing the scene via an infrared (IR) light source and cameras are augmented to perceive IR as well as visual spectra. Very recently, the D400 family [53] (including RealSense D435) was released, which features long-range capabilities and high depth resolution with wide field of view, and therefore has the potential to deliver more accurate depth maps compared to RealSense R200. The built-in stereo algorithm in these cameras uses a handcrafted binary descriptor in conjunction with a semi-global matching scheme [52]. Smart Glasses and D435 benefit from the global shutter in fast-moving and outdoor applications. They provide a reasonable robust solution in both indoor and outdoor scenarios.

Comparatively, the pRGB-D Sensor (ZED integrated) and ZED Mini implement GPU-accelerated global optimization algorithms to attain dense and large-scale depth perception at distances more than 10m. Although pRGB-D Sensor and ZED Mini can also be used indoors, they are more reliable in richly-textured outdoor environments with abundant local correspondences. Overall, 3D imagery based on these RGB-D sensors has been a de-facto standard in diverse robotic vision tasks and the driving force behind the revolution of many algorithms. In this work, these four RGB-D sensor systems are selected mainly taking into the account the environmental adaptability and portability.

The key specifications for navigation assistance of these perception systems are summarized in Table 1. Although these systems can support larger resolutions (*e.g.*, 1920×1080 or 1280×720), we set their output resolutions close to the VGA resolution for streaming to facilitate fair comparison, and use equal or smaller resolution for efficient segmentation, *e.g.*, 320×240. Notably, RealSense D435 can deliver depth maps covering a wide field of view at 91.2°×65.5°, but the field of view of its RGB camera is relatively limited (69.4°×42.6°), such that we project the depth measurements into the RGB images to obtain aligned

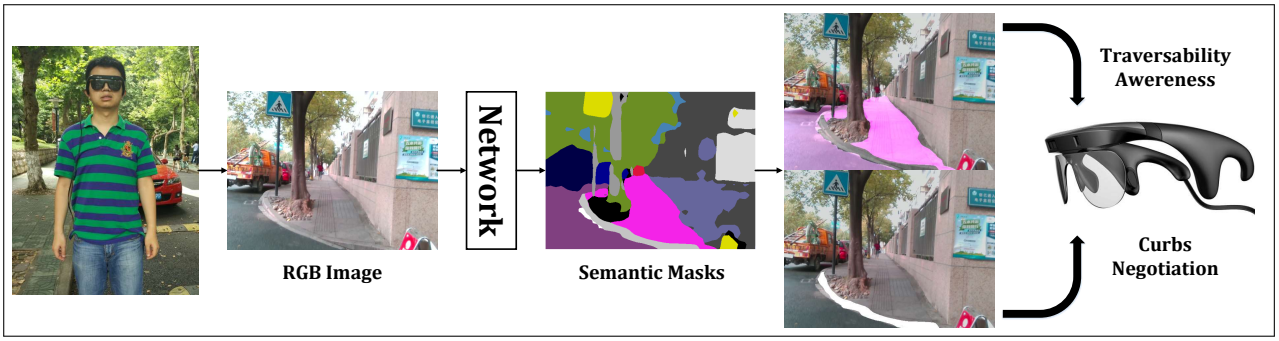

Fig. 2. Overview of a wearable navigation assistance system with a semantic perception framework supporting traversability awareness and curbs negotiation.

per-pixel RGB-D information. Accordingly, ZED Mini attains the largest field of view with aligned RGB-D information. The minimum depth range is determined by multiple factors including field of view, disparity searching range, and possible overexposure caused by the projected speckles [55]. A small minimum range denotes that the system can deliver depth information of a very close obstacle, which is important for safety-critical navigation assistance systems [56]. As far as the maximum values are regarded, there exists no threshold for Smart Glasses and RealSense D435 because the maximum depth range of active stereo-based sensors varies depending on the scene and light conditions. Nevertheless, we notice their depth maps become less dense and effective at distances over 10m. Regarding the size, RealSense D435 and ZED Mini are both miniaturized but the thickness may be a burden which is induced by the lens design to attain large field of view. In contrast, RealSense R200 is very suitable for being integrated in a wearable device due to its small size and light weight (65g), and it has been already prototyped as the form of a pair of Smart Glasses, which weighs 117g after integration. It is worth noting that in the RealSense D400 family, smaller depth modules are available (*e.g.*, D430), but the power consumption (3.5W) is a bit high for wearable devices with its laser speckle emitter projecting a random focused dot pattern.

C. Semantic Segmentation Architecture

To address the deployability of per-pixel semantic scene parsing, our architecture is designed according to the encoder-decoder networks like SegNet [27], ENet [28] and our previous ERFNet [30], as expanded in Figure 3. Table 2 also gives detailed description of the proposed network built on spatial factorizations, hierarchical dilations and pyramidal representations, with central consideration of efficiency and robustness. PyTorch, TensorFlow implementations and NVIDIA TX1/TX2 deployment codes corresponding to our SS framework have been open-sourced at [57], [58] and [59] respectively.

C.1. Spatial Factorization

Current trends have paved the road to high SS accuracies since the appearance of residual layers [60] that avoid the degradation problem, allowing the gradient to be propagated through a large number of layers, thus the network will be directed towards learning the residual representation on identity mapping. The rationale behind is that identity mapping with shortcuts can facilitate the optimization of deep networks, since it iteratively generates small magnitudes of responses by passing main information layer by layer. Basically, the residual layer adopted in state-of-the-art networks has two instances, the bottleneck version and the non-bottleneck design. Utilizing 1D factorizations of the convolution kernels, "Non-bottleneck-1D" (Non-bt-1D) was redesigned in our previous work [30] to rationally strike a balance between the efficiency of bottleneck and the learning capacity of non-bottleneck. The spatial factorization into separable asymmetric convolutions enables an efficient use of residual layers to extract feature maps and infer semantic predictions in real time.

C.2. Hierarchical Dilation

Starting from the observation that increased number of layers help to learn more complex and abstract features, which leads to increased SS accuracy but also increased running time, we propose the Hierarchical Dilated Non-bottleneck-1D block (HD-1D block), which has two instances including the 4×2 hierarchical architecture and the 3×3 hierarchical design as illustrated in Figure 4. Compared with conventional schemes [10, 36], the proposed block is composed of multilevel parallel dilated factorized convolutions with various dilation rates. This hierarchical structure enables the network to capture large Field-of-View (FoV) for varying object sizes and reduce susceptibility to overfitting on existing datasets, while the enlarged receptive field is earned with less increased depth of deep CNNs. Vitally, the bypass connection extends the proposed HD-1D block from a straightforward repeated parallel structure by allowing each dilated layer to attain access to other Non-bt-1D layers, which positively leads to an implicit deep supervision, such that the depth of CNNs is not completely sacrificed. In this sense, our HD-1D block offers context assimilation on large FoV, inference speedup and competitive accuracy compared with the original architecture that sequentially stacks dilated Non-bt-1D layers [30].

C.3. Pyramidal Representation

Another key reconstruction from the backbone architecture (initial efficient residual factorized network) lies in the substitution of the original decoder with the pyramidal module of the PSPNet [61], which presumably exploits better the contextual priors from such representation by combining semantic cues from different sub-regions of the feature map. This modification also offers critical insights into wearable navigation assistance [5], including effectiveness to learn common knowledge, robustness to object size, and smoothness of representation.

In pursuit of these specific characteristics, the pyramid pooling module introduced by PSPNet is leveraged to harvest different sub-region representations, followed by up-sampling and concatenation layers to form the final feature representations.

Table 1. Specifications of the perception systems.

	Smart Glasses [1]	pRGB-D Sensor [20]	RealSense D435 [53]	ZED Mini [54]
Horizontal field of view	55.1°	84.9°	69.4°	**87.3°**
Vertical field of view	42.7°	54.5°	42.6°	**56.5°**
Resolution	640×480	640×360	640×360	640×360
Minimum depth range	350mm	700mm	250mm	**200mm**
Maximum depth range	Approx. 10m	**20m**	Approx. 10m	12m
Baseline	70mm	120mm	50mm	63mm
Size	**101.6mm×9.6mm×3.8mm**	175mm×30mm×33mm	90mm×25mm×25mm	124.5mm×30.5mm×26.5mm
Weight	117g (65g)	159g	100g	**62.9g**
Power	**1.6W**	1.9W	3.5W	1.9W

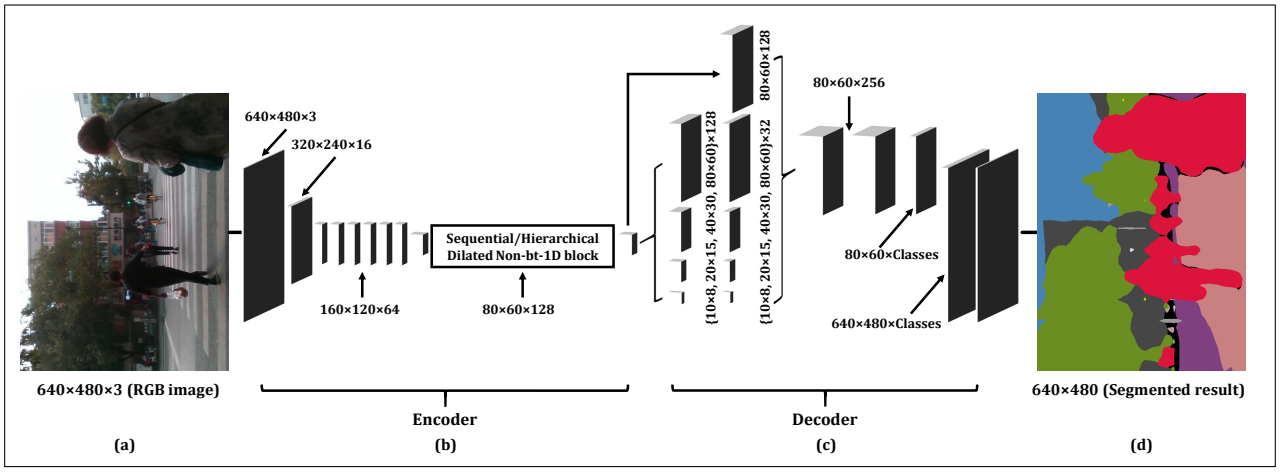

Fig. 3. The proposed architecture. (a) Input, (b) Encoder, (c) Decoder, (d) Prediction.

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 pyramidal module contains the feature map from the encoder with varied sizes. As shown in Figure 3(c), to maintain the weight of global feature, namely the feature map with a channel number of 128 outputted by the encoder, a 1×1 point-wise convolution layer is appended after each pyramid level to 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 is Figure 3(c), the level size N equals to 4 and we decrease the number of feature maps from 128 to 32. To be brief, the module uses a global pooling layer and 3 finer non-overlapping pooling layers with 4 different bin sizes. Subsequently, these low-dimension feature maps are directly up-sampled to obtain the same-size features as the original feature map through bilinear interpolation for smoothness consideration. Figure 3 showcases the overall perspective of the architecture, depicting the feature maps generated by each of the block, consecutively from the RGB input to the per-pixel class probabilities and final prediction for the real-world urban scenario observed by Smart Glasses.

D. Data Augmentation Methods

In essence, deep CNN architectures have evolved a high reliance on the training data, since the features directly learned by a model depend entirely on the images that are fed during this process. In this regard, the diversity of data plays an essential role to obtain trained models with good generalization capacity. On the one hand, CNNs need to learn the broad variety of patterns, and enable the filters to discriminate well between semantic categories. Besides, CNNs need to be prevented from learning irrelevant features. Data augmentation can fulfill exactly these purposes.

In this framework, we apply a vast array of methods with the aim of augmenting the limited set of data to improve the robustness [62]. Among these techniques, some have an effect on the geometry of the categories (i.e., position and shape) and others have an effect in the texture (i.e., illumination and color). Regardless of geometry or texture data augmentation techniques used during training, the CNN will be affected in learning patterns from the datasets in order to produce SS output. In this sense, it is crucial to augment both to improve the network performance in unseen domains and robustify SS across RGB-D cameras.

D.1. Geometric Augmentations

Horizontal flipping is performed at a 50% chance to add invariance to orientation. Translation augmentation, aspect ratio augmentation are enabled together with scaling and cropping, by performing random rescaling uniformly 0.5 and 1.0 times the original height of the image size and another random value to the image width, and combining it with randomly cropped regions of the image to keep the same resolution in the training batch. These three augmentations prevent the CNN from seeing always the same position of the training images, so it doesn't always generate the same activations from the very first layer.

Table 2. 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	Original RGB image	3	640×480
	1	Down-sampler block	16	320×240
	2	Down-sampler block	64	160×120
	3-7	5×Non-bt-1D	64	80×60
	8	Down-sampler block	128	80×60
	9-16/17	Dilated Non-bt-1D layers	128	80×60
DECODER	0	Original feature map	128	80×60
	1	Pooling and convolution	32	80×60
	2	Pooling and convolution	32	40×30
	3	Pooling and convolution	32	20×15
	4	Pooling and convolution	32	10×8
	5	Up-sampler and concatenation	256	80×60
	6	Convolution	C	80×60
	7	Up-sampler	C	640×480

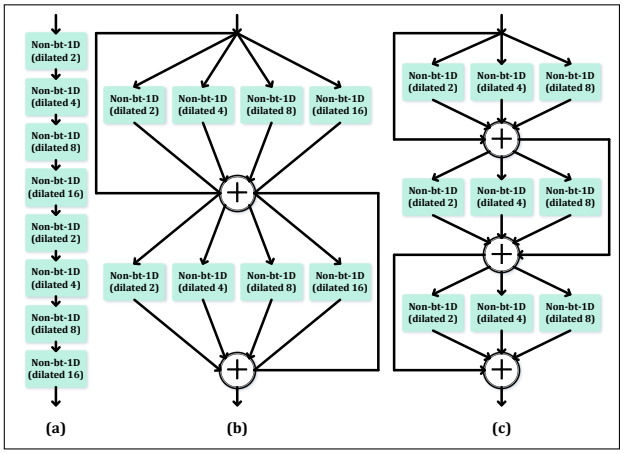

Fig. 4. Sequential and hierarchical architectures of dilated Non-bottleneck-1D (Non-bt-1D) layers. From left to right: (a) Sequential architecture, (b) 4×2 hierarchical architecture, (c) 3×3 hierarchical architecture.

In addition, invariance can be reached against diverse aspect ratios (e.g., 4:3, 16:9) that can be specific to each camera in a navigation assistance system. Motivated by the fact that objects might appear with angle variances in the scene observed by the wearable devices, we implement the rotation without cropping by sampling distributions from the angles [-20°, 20°].

D.2. Textual Augmentations

Firstly, we alter brightness following a uniform distribution between [0.8, 1.2] to improve the illumination invariance. Secondly, we augment contrast by uniformly choosing a jittering value from the range [0.8, 1.2] to aggrandize invariance to dynamic range of the scene and the camera. Thirdly, we augment saturation within a uniform distribution between [0.8, 1.2] to add invariance to different camera sensitiveness in capturing color, and perform hue augmentation by adding a value from the range [-0.2, 0.2] to the hue value channel of the HSV representation, to attain invariance against color deviations. Last but not the least, sharpness is the term related to the edge contrast, the texture richness and density, as well as the clarity of details (e.g., pixel-exact segmentation of curbs versus traversable areas). Augmenting the sharpness-diversity of the fed images helps the network to retain invariance to the blurriness due to diffused light or camera shake which may vary across navigation assistance systems. Otherwise, the lack of invariance may corrupt the prediction especially around boundaries, leading to segmentation inconsistency. We alter sharpness within a uniform distribution between [0.9, 1.1].

4. EXPERIMENTS AND DISCUSSIONS

A. Experiment Setup

The real-world experiments are performed in public spaces around the Polytechnic School at University of Alcalá (Madrid) during a typical sunny workday in Spain, where the shadows and complex sunlight situations can be challenging to the SS frameworks. We captured the Campus and the surrounding Metropolitan scenes using four different perception systems. For the wearable navigation assistance systems including KR-Vision Smart Glasses and pRGB-D Sensor, the volunteers have worn these prototypes along the predetermined route. For the portable sensors including RealSense D435 and ZED Mini, the volunteers held them at the height of head-worn devices to capture the egocentric vision datasets. Each prototype has been carried along the same route around the campus (about a 1000-m trajectory). Although the cross-system images are not strictly aligned, they offer the comparable, ordinal data in our robustness study.

Our raw dataset contains over 6500 images sub-sampled at 4Hz. Each perception system has captured around 1600-1700 RGB-D images, out of which 100 evenly distributed are finely annotated. In total, 400 RGB-D images are collected and manually labeled with pixel-wise SS ground-truth for traversability-related categories: roadways, sidewalks and curbs. This RGB-D-SS dataset has been offered to the community, publicly available at [63]. Similar to RANUS [9], our RGB-D-SS is also a multi-sensorial street scene dataset. Compared with state-of-the-art evaluation-oriented datasets like WildDash [49] (70 public test cases), DarkZurich [46] (20 fully labeled images) and PASS [8] (400 annotated panoramas), our RGB-D-SS is large enough and it specifically features cross-system RGB-D observations.

The metrics used in this paper correspond to Intersection-over-Union (*IoU*) and Pixel-wise Accuracy (*PA*) that are prevailing in semantic segmentation challenges:

$$IoU = \frac{TP}{TP + FP + FN} \quad (1)$$

$$PA = \frac{CCP}{LP} \quad (2)$$

where *TP*, *FP*, *FN* are respectively the number of True Positives, False Positives and False Negatives at the pixel level, where *CCP* and *LP* denote the number of Correctly-Classified Pixels and Labeled Pixels, respectively. We also use mean IoU (*mIoU*), mean PA (*mPA*), Pixel-wise Accuracy of traversable area parsing (*tPA*) that is calculated by merging the roadways, sidewalks and other traversable scene categories as traversable areas. Mean Value (*MV*) of pixel-wise accuracy and Coefficient of Variation (*CV*) across navigation assistance systems, are also employed to analyze the real-world performance:

$$MV = \frac{1}{n}\sum_{i=1}^{n} PA_i \quad (3)$$

$$CV = \frac{\sqrt{\sum_{i=1}^{n}(PA_i - MV)^2/n}}{MV} \quad (4)$$

B. Training Setup

As far as navigational semantic segmentation is concerned, the challenging Mapillary VISTAS dataset [7] is chosen as the training dataset, because it consists of many traversability-related classes across a wide spectrum of outdoor street-level scenes from different roadways, sidewalks (pavements), complex intersections and roundabouts, which are typical scenarios during assisted navigation. In addition, it attains vast geographic coverage, containing images from different continents. This is important to enhance reliability because traversability-related objects like curbs are not exactly the same in different streets, campuses and countries. In total, we have 18000 images for training and 2000 images for validation with pixel-exact annotations. Regarding the assistive awareness of semantics of interest to be rendered for users, 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. The images come from VISTAS dataset have very large resolution (higher than 1920×1080), but they are homogenized to a resolution of 640×480 to be used for random cropping during the data augmentation phase.

For our deep neural networks, Adam optimization [64] is employed for model training, which is initiated with a learning rate of 5×10^{-5} that decreases exponentially across epochs, operated with a batch size of 15, momentum of 0.9 and weight decay of 2×10^{-4}. The maximum epoch number is 300 for all models trained with data augmentations, separated into two stages: first the encoder is trained by mapping an input image to a down-sampled label; then we append corresponding decoder to the trained network followed by a pixel-wise classifier. Notably, focal loss [65] is adopted as the criterion during training until it converges:

$$Focal_{loss} = \sum_{i=1}^{W}\sum_{j=1}^{H}\sum_{n=0}^{N}(1-\mathbf{P}_{(i,j,n)})^2\mathbf{L}_{(i,j,n)}log(\mathbf{P}_{(i,j,n)}), \quad (5)$$

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). In contrast, it suppressed lightly the loss contribution of wrongly-segmented pixels ($\mathbf{P}_{(i,j,n)}$=0.1, $(1-\mathbf{P}_{(i,j,n)})^2$ =0.81). In this way, the focal loss concentrates the network training on wrongly-segmented pixels or hard pixels [66]. Under the supervision, we found models can yield better results than conventional cross-entropy loss on VISTAS dataset, as it contains some less-frequent yet important classes such as traffic lights and hazardous curbs.

C. Real-Time Performance

As displayed in Table 3, the frame rates of our sequential/hierarchical ERF-PSPNets are tested. There are some top-performing networks such as DeepLab V3+ [12] and Dense Relation Net [13] but they involve computationally-intensive models, which are expensive to train on low-cost GPUs, and too heavy to deploy on embedded GPUs or wearable devices. There are also some models like FRRN [11] and ICNet [31] which rely on high-resolution input that is normally kept for redundancy to assist cognition and navigation of the visually impaired. ERF-PSPNets differ from those complex networks in terms of efficiency and application scenario. Accordingly, we contrast our architectures with two well-known state-of-the-art networks for real-time semantic segmentation including ENet [28] and LinkNet [29]. At 320×240, a resolution that is enough to recognize any urban scene accurately and create effective feedback for navigation assistance, our 4×2 hierarchical architecture is the fastest when testing on a cost-effective processor with a single GPU GTX1050Ti. Admittedly, the runtime of LinkNet is not able to be tested due to the inconsistent tensor sizes at down-sampling layers. For this reason, we test at 448×256, another efficient resolution at which most of the architectures can be evaluated, where our 4×2 hierarchical architecture is also the fastest, outperforming LinkNet by a slight margin. At 640×480, the VGA resolution, ENet is the fastest, while our models still maintain near real-time prediction. This result verifies that the speed analysis of semantic segmenters should not be confined at single size since the rising tendencies of runtime with regard to image resolution are different for various end-to-end networks. Nevertheless, for wearable navigation assistance, 320×240 is arguably the optimum resolution of the three resolutions, since pixel-exact features are less desired by the blind user, but entail higher input resolution that incurs longer processing latency. Still, the mean IoU values of our models tested on VISTAS dataset [7] are significantly higher than ENet and LinkNet. Here, ENet and our sequential/hierarchical ERF-PSPNets are trained at 320×240, while LinkNet is trained at 448×256, both with the full set of data augmentations.

Table 3. Speed and semantic segmentation accuracy analysis. "FR": Frame Rate on a cost-effective GPU GTX1050Ti, "mIoU": mean Intersection-over-Union.

Network	FR at 320×240	FR at 448×256	FR at 640×480	mIoU
ENet [28]	66.2FPS	57.5FPS	**41.8FPS**	33.6%
LinkNet [29]	N/A	72.5FPS	31.6FPS	39.4%
Sequential ERF-PSPNet	75.8FPS	62.5FPS	29.1FPS	**48.4%**
4×2 Hierarchical ERF-PSPNet	**82.0FPS**	**73.0FPS**	33.9FPS	47.1%
3×3 Hierarchical ERF-PSPNet	77.5FPS	69.4FPS	32.2FPS	48.1%

For the sake of completeness, we also test on an embedded GPU Tegra TX1 (Jetson TX1) that enables higher portability of navigation assistance systems, while consuming less than 10 Watts at full load. At 320×240, and our models achieve more than 22.0FPS. Evidently, the hierarchical architectures are both faster than the sequential architecture while only causing minor drop in segmentation performance.

D. Segmentation Accuracy

Table 4 details the accuracy of 17 main navigational classes and the mean IoU values. It could be told that the accuracy of most classes obtained with the proposed ERF-PSPNet exceed the existing architectures that are also designed for real-time applications. Our architecture has the ability to collect rich contextual information without major sacrifices of learning from textures. Accordingly, only the accuracy of Sky is slightly lower than LinkNet, while most important classes for traversability and traffic safety perception are both higher, given that these networks are both

trained with the same data augmentation strategy. Noticeably, our ERF-PSPNets outperform ENet and LinkNet by a large margin on less frequent classes, such as Motorcycle, Bike lane, Curb and Rider, which implies that the ability to gather contextual information is critical especially for those classes, because it is generally difficult to classify these classes by using textures from few training data. This finding is also consistent with [23].

When comparing the hierarchical ERF-PSPNets with the sequential version, although the mean IoU values of all classes used for training is lower, they offer some benefits. Firstly, in spite of being a possible subjective measure, the mean IoU values of 17 main navigational classes are higher than that achieved with the sequential design. Secondly, on some semantic classes of interest, hierarchical design achieves best accuracies. For example, 4×2 hierarchical ERF-PSPNet outperforms the sequential version on Motorcycle by a margin of 5.6%, and 3×3 hierarchical ERF-PSPNet surpasses the sequential version on Bike lane by a margin of 6.9%. We believe such remarkable differences are not caused by the random process of network training. As indicated by the statistics in Figure 5(b), Bike lane and Motorcycle correspond to the least frequent classes. For Rider, the class with the third lowest instance frequency, 3×3 hierarchical design also exceeds the sequential one, and the accuracy difference between 3×3 and 4×2 versions reaches up to 6.8%. It suggests that the proposed HD-1D block is promising to boost the segmentation performance on less frequent classes. Following the rationale that the ability to capture contextual information helps to reach better scores on classes with few training data, this result makes sense since the hierarchical structure is beneficial to capture objects with varying sizes, which is a critical aspect of context information [61]. Another evidence is that ENet fails to classify Motorcycle well as its accuracy is only 0.1%. Such problem of ENet was also reported in [36]. Noticeably, the performance gap is more related with instance frequency rather than pixel frequency as illustrated in Figure 5.

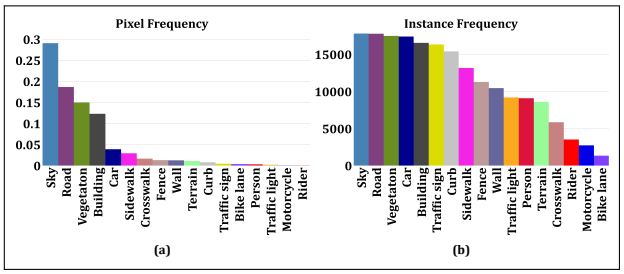

Fig. 5. Class frequency of the VISTAS dataset: (a) Pixel frequency (portion of labeled pixels), (b) Instance frequency (number of images with at least one labeled instance).

It is worthwhile to mention that we have also pretrained the sequential version of ERF-PSPNet on ImageNet [67], eventually the mean IoU value reaches 48.8%, which is marginally higher than that achieved using the two-stage "from scratch" strategy (48.4%). This result reveals that although the transferability of features of pretraining on a large dataset is advantageous, our models can also reach good accuracy when trained on a single dataset without the need of pretraining that adds training complexity and may suppose commercial limitations. For the pretrained sequential ERF-PSPNet without data augmentations, most *IoU* values are lower than that trained with data augmentations. The *mIoU* gap between 48.8% and 46.4% demonstrates that, although the performed data augmentations were dedicated to robustifying semantic perception against diverse image styles seen by navigation assistance systems in the real world, they also boost the performance on unseen validation data.

E. Real-World Performance

Taking an essential step further than experimenting on VISTAS benchmark, the real-world performance of these trained networks are analyzed on our RGB-D-SS dataset that is captured using four perception systems. We use Pixel-wise Accuracy (*PA*) to eliminate the influence of unlabeled classes and facilitate fair comparison between CNN-based approaches and depth-based approaches, since traditional traversability perception algorithms have no ability of direct classification but only detect the ground plane. Admittedly, our real-world dataset is created by only annotating pixel-wise ground truth for traversability-related semantic classes, so it would be reasonable to focus on the study of robustness using accuracy-based metrics as displayed in Table 5.

ENet is well known as an extremely efficient network that sacrifices its model size and learning capacity to implement real-time semantic prediction. When comparing the real-world performance between different networks, we notice that although LinkNet achieves higher accuracy values than ENet on VISTAS dataset, it is not able to exceed ENet on all classes in the real-world setting, especially the sidewalks. This reveals that LinkNet still suffers from limited learning capacity, which may bias the appearances of scenes to be analyzed. Additionally, both *CV* metrics calculated for the accuracy terms with LinkNet are higher than ENet. This comparison indicates that using LinkNet results in a larger performance variance across navigation assistance systems, which has preliminarily illustrated that "accuracy" and "robustness" are not equivalent concepts when it comes to real-world performance.

Regarding our models, the sequential ERF-PSPNet outperforms other networks including ENet and LinkNet. It achieves the best accuracy with pRGB-D Sensor, RealSense D435 and ZED Mini in terms of *mPA* and *tPA*, despite the sub-optimal performance second to a certain hierarchical version with Smart Glasses. Crucially, the effectiveness of data augmentations on the issue of real-world cross-system robustness is evident from the experimental results, which are calculated using the outputs that are produced by models with or without data augmentations (refer to Table 5 for "Pretrained Seq (Augmented)" versus "Pretrained Seq (No Augs)". On the one hand, almost all real-world segmentation accuracy values have been boosted, which means that with the augmented model in such unseen domain, the predicted segmentation map can be more reliable and most traversable areas (over 94% at the pixel level) can be successfully detected. On the other hand, both *CV* values have been decreased by a large extent, which further demonstrates that the accuracy gaps between RGB-D sensor systems have been reduced. In summary, the two-level evidence indicates that the semantic traversability perception approach has been significantly robustified.

We have also collected the accuracy data (*tPA* values) of depth-based segmentation approaches including 3D-RANSAC-F [15] and FreeSpaceParse [19], by projecting the 3D information into the RGB image view for pixel-aligned comparison. For Smart Glasses, since the original depth map delivered by RealSense R200 is sparse and noisy, we have implemented a guided filter [16, 68] to eliminate small segments, fill invalid holes and enhance the depth image. Otherwise, the depth-based approaches will yield ill-posed detection results. For other per-

Table 4. Accuracy analysis using Intersection-over-Union (IoU).
"Mean-17": mean IoU value of 17 navigation-related classes, "Mean-27": mean IoU value of all 27 classes used for training.

Network	Traffic light	Car	Motorcycle	Traffic sign	Road	Sidewalk	Bike lane	Curb	Fence	Wall	Building	Person	Rider	Sky	Vegetation	Terrain	Crosswalk	Mean-27	Mean-17
ENet [28]	25.0%	71.2%	0.1%	39.2%	82.5%	57.2%	12.4%	33.0%	27.8%	35.1%	76.0%	32.6%	2.7%	96.4%	81.1%	52.9%	51.0%	33.6%	45.6%
LinkNet [29]	34.6%	74.4%	20.6%	45.1%	84.0%	58.2%	19.7%	37.1%	33.5%	37.7%	78.2%	42.3%	16.2%	**97.2%**	83.3%	54.9%	51.9%	39.4%	51.1%
Seq ERF-PSPNet	**38.2%**	**76.4%**	36.5%	51.9%	85.6%	63.8%	30.5%	**43.1%**	41.6%	47.2%	**80.6%**	**48.1%**	40.4%	96.6%	**83.9%**	**59.6%**	**59.1%**	**48.4%**	57.3%
4×2 Hier ERF-PSPNet	35.6%	76.1%	**42.1%**	50.7%	85.2%	62.1%	32.8%	**43.1%**	40.9%	48.0%	80.0%	47.1%	34.1%	96.5%	83.6%	59.3%	57.8%	47.1%	57.8%
3×3 Hier ERF-PSPNet	36.6%	76.2%	36.3%	**52.0%**	**85.8%**	**64.5%**	**37.4%**	42.7%	**41.9%**	**49.8%**	80.4%	47.1%	**40.9%**	96.5%	83.8%	58.9%	58.6%	48.1%	**58.2%**
Pretrained Seq	37.1%	**75.9%**	**39.8%**	**52.8%**	**85.9%**	**65.1%**	**36.0%**	**42.9%**	**42.9%**	**47.6%**	**80.5%**	**49.9%**	**40.5%**	**96.5%**	**84.1%**	**60.1%**	60.0%	**48.8%**	**58.7%**
Without Augs	**38.2%**	75.0%	31.7%	50.5%	85.2%	61.7%	30.0%	**42.9%**	40.1%	45.1%	79.9%	49.0%	35.8%	**96.5%**	83.9%	58.8%	**60.6%**	46.4%	56.8%

Table 5. Real-World robustness analysis using Pixel-Wise Accuracy (PA).
"mPA": Mean Pixel-wise Accuracy for segmentation of road, sidewalk and curb, "tPA": Pixel-wise Accuracy for segmentation of traversable areas merged from road, sidewalk and traversable classes, "CV": Coefficient of Variation.

Approach	Metric	Smart Glasses [1]	pRGB-D Sensor [20]	RealSense D435 [53]	ZED Mini [54]	CV
ENet [28] (Augmented)	mPA	57.3%	72.3%	70.1%	70.1%	8.8%
	tPA	91.8%	93.9%	93.0%	93.0%	**0.8%**
LinkNet [29] (Augmented)	mPA	55.5%	76.4%	70.0%	73.2%	11.6%
	tPA	92.6%	95.3%	94.3%	94.8%	1.1%
Seq (Augmented)	mPA	69.5%	**80.8%**	**80.0%**	**79.6%**	5.9%
	tPA	96.9%	**97.0%**	**96.2%**	**96.4%**	1.2%
Hier 4×2 (Augmented)	mPA	**73.8%**	78.3%	75.2%	73.8%	**2.4%**
	tPA	96.8%	93.7%	92.8%	90.4%	2.5%
Hier 3×3 (Augmented)	mPA	72.7%	79.3%	78.9%	72.4%	4.3%
	tPA	**97.6%**	93.1%	94.7%	88.1%	3.7%
Pretrained Seq (Augmented)	mPA	**78.1%**	**81.4%**	**77.4%**	**77.4%**	**2.1%**
	tPA	**95.4%**	**94.5%**	**95.5%**	**94.1%**	**0.6%**
Pretrained Seq (No Augs)	mPA	60.8%	79.1%	69.5%	75.3%	9.7%
	tPA	88.7%	94.3%	91.7%	92.6%	2.2%
3D-RANSAC-F [15] (depth-based)	tPA	70.6%	84.0%	92.6%	90.1%	10.1%
FreeSpaceParse [19] (depth-based)	tPA	87.1%	66.5%	86.3%	80.9%	10.3%

ception systems, we directly use the original depth map without any post-processing. It turns out that *CV* values obtained with depth-based approaches are both higher 10.0%, which are higher than all *CV* values obtained with CNN-based approaches (see those computed from *tPA* values across systems in Table 5). This has numerically validated the generalization capacity of SS-based traversability detection framework. Based on our study, the research community should be motivated to revolutionize traditional RGB-D sensory assistive technology by applying deep learning algorithms for unification and generalization considerations.

F. Robustness to Changes of Sensorial Factors

Although our purpose is to robustify semantic perception and shrink the performance gap between different systems, one frequently-asked question is that which RGB-D camera supports the highest accuracy. To answer this question, we have calculated the pixel-wise accuracy at different ranges taking into account that short-range of ground area detection helps to determine the most walkable direction [20, 40], while superior path planning could be supported by longer traversability awareness [16]. Since the minimum detectable depth ranges of these RGB-D sensors are different as discussed in Section 3B, the experimental data are collected within 9 ranges: 1-2m, 2-3m, ..., 9-10m. As visualized in Figure 6 regarding the pixel-wise accuracy of the critical traversable area parsing and the overall accuracy recorded in Table 5, the semantic perception framework yields the highest accuracy for traversability awareness with the pRGB-D Sensor in a dominant part of situations, which is also consistent on specific classes including roadways, sidewalks and curbs.

We are very curious about why pRGB-D Sensor delivers the best performance. Generally, the accuracy will be highly related to the domain style of images used for training. This might be one reason, but the pRGB-D Sensor integrated ZED, while ZED and ZED Mini share similar image style yet distinct performance. More rationally speaking, our first observation is that the pRGB-D Sensor has the better optical imaging performance with less stray light as displayed in Figure 7 (see the sky region). The second hypothesis is regarding images from VISTAS dataset [7], the distribution of the focal length is mostly concentrated in the 25-35mm range and the peak is around 29mm (35mm equivalent). The focal lengths of pRGB-D Sensor (ZED integrated), ZED Mini and RealSense D435 are 27.9mm, 26.8mm and 19.3mm, respectively. Accordingly, there is nearly an order of magnitude more images involved in training stage which shares the similar focal

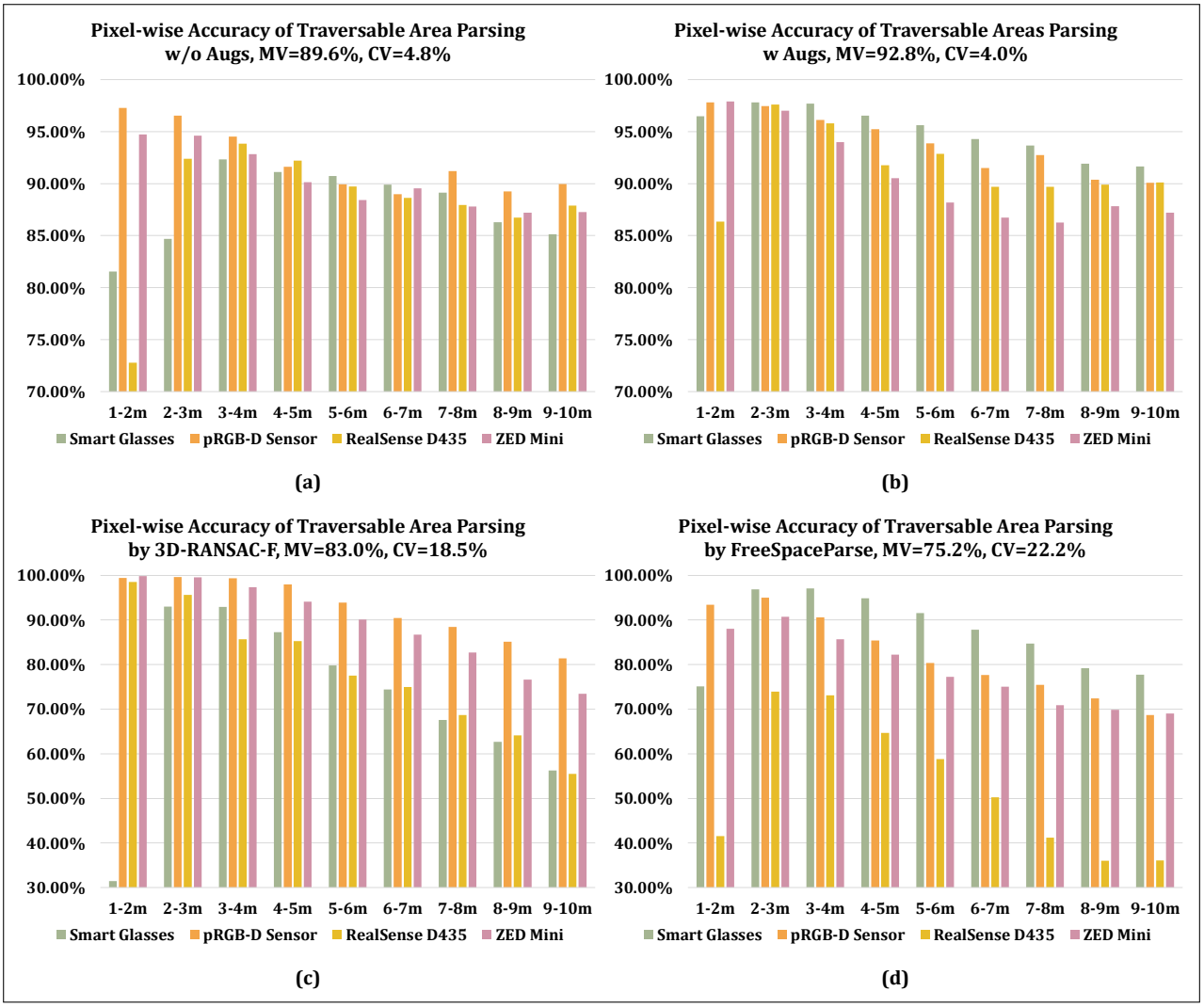

Fig. 6. Comparison of the pixel-wise accuracy values of traversable area parsing at different ranges across perception systems produced by (a) our sequential ERF-PSP model trained without data augmentations, (b) our model trained with all data augmentations, (c) 3D-RANSAC-F [15] and (d) FreeSpaceParse [19].

length to the images captured by a ZED-based prototype (pRGB-D Sensor), than ZED Mini or RealSense D435. This indicates that even though a group of geometric data augmentations have already be performed, the unaltered focal length still has a great influence on semantic prediction. The underlying rationale is that, contextual information of traversability-related categories should be more consistent in 3D space, while 3D-to-2D imaging process must satisfy some strict projective relationship. For example, the image of an object projected by a long-focal-length camera in the far distance could be exactly the same as the one captured by a short-focal-length at a short distance [69]. Such indistinguishability may bias the network in predicting semantics of traversability. The essential role of data augmentations regarding focal length have also been investigated in the research line of fish-eye image segmentation [22]. In this sense, future efforts would be dedicated to chaining scene depth inference and embedding camera's intrinsic parameters such as focal length in the semantics/depth prediction models. For Smart Glasses, the major problem is that it suffers from motion blur, rolling shutter artifacts and bluish color deviation as shown in Figure 7. Still, after applying a batch of data augmentations, the performance boosts of SS with Smart Glasses can be clearly observed in Figure 6. The main insight gained from our experiments is that in essence, the gap between the concepts of "accuracy" and "robustness" is not only a matter of training images or CNN learning capacity, but also a matter of data diversity and optical characteristic.

More importantly, it is consistent that with data augmentations, the *MV* metric at the detectable ranges of these systems has been improved as illustrated in Figure 6(a)(b), which demonstrates that the robustness has been enhanced in the real world. Especially, *CV* metric can be used to describe the performance variation across systems beyond general robustness. Similarly, by applying data augmentations, all *CV* values have been decreased regarding the pixel-scale segmentation of specific classes including roadways, sidewalks and curbs. When comparing the depth-based approaches with CNN-based approaches (see Figure 6(a)(b) versus Figure 6(c)(d)), we reach the same conclusions as drawn in Section 4E: CNN-based approaches are significantly more robust than depth-based approaches if they should be deployed in various navigation assistance systems. In addition to the robustness aspect, an issue we want to report is that no-

ticeably, RealSense D435 suffers from low detection range and accuracy using FreeSpaceParse as manifested in Figure 6(d). This is due to the less reliable depth map at close ranges. One representative example can be found along the fifth row in Figure 7(c)(f).

Figure 7 exhibits the montage of pixel-wise masks generated by our approache, ENet, FreeSpaceParse and 3D-RANSAC-F. Qualitatively, our approach not only yields more smooth, more consistent and more complete segmentation at both close ranges and long ranges, but also retains the outstanding ability to perceive hazardous curbs within this unified perception framework. In summary, semantic cognition of traversability has been enhanced with our augmented model, leading to the improvement of accuracy and robustness on both numerical metrics and visualization results.

G. Robustness to Changes of Environmental Conditions

To elaborate the effectiveness of data augmentations on the robustness to environmental factors, we present a set of comparisons of segmentation results for some extreme cases in our environment in Figure 8. For example, the water hazards on the road, the overexposure in the scenarios with high dynamic range of illumination level, the unseen domain with a retrofitted car, and the shadows can be very challenging for robust semantic perception. In Figure 9, we showcase cross-continent results for data recorded not only around the Polytechnic School at University of Alcalá in Madrid, but also in public spaces around Westlake, the Yuquan Campus, the City College at Zhejiang University, and Holley Metering Campus in Hangzhou. Diverse segmentation masks with different illumination conditions (cloudy vs. sunny), different aspect ratios (4:3 vs. 16:9) and different viewpoints (from roadways, sidewalks and off-roads) have been combined and visualized. In summary, it can be seen in all qualitative examples, our model delivers impressive segmentation results even under challenging conditions thanks to its inherent robustness and the extremely positive effect of data augmentations in robustifying semantic perception across all kinds of domains and RGB-D observations.

5. CONCLUSION AND FUTURE WORK

Semantic Segmentation (SS) stands out as an effective approach to unify different detection tasks. This paper considerably extends the previous field navigation-dedicated work [5] by covering the detection of curbs and traversable areas in a semantic traversability perception framework. We have systematically studied the robustness of Convolutional Neural Network (CNN)-based semantic segmenters. After designing a cluster of efficient deep architectures, data augmentations have been combined to enhance the model such that it will be ready to produce accurate segmentation across real-world navigation assistance domains.

In summary, the sequential ERF-PSPNet achieves the highest segmentation accuracy while the 3×3 hierarchical design reaches the fastest speed at efficient resolutions for wearable semantic cognition. Regarding the cameras, the customized pRGB-D Sensor helps to attain the best performance of traversability awareness in a dominant part of conditions and ranges, while the pair of Smart Glasses offers an optimal light-weight solution for wearable navigation assistance.

In our research line, the unification and generalization capacity of SS-based semantic perception have been validated on numerical metrics and visualization results. Based on this comparative study, entrepreneurs and roboticists can answer justifiably that the robustness to diverse environmental and sensorial factors is reachable, while researchers and engineers should be encouraged to apply deep learning-based segmenters in upper-level applications assisting semantic cognition of robotics and vulnerable road users.

Future work will be contributed to the continuous improvement of our wearable navigation assistive framework towards long-term autonomy. We will resort to relevance-aware loss functions [33, 70] to boost the segmentation performance on less frequent yet safety-critical semantic categories. Another promising direction is to include depth-wise separable convolution as an additional shallow branch for spatial fine-tuning [34] to extend the use of our networks to even wider variety of application scenarios. We also have the intention to adapt the semantic cognition to RGB-D input, as well as omni-directional imagery by using a Panoramic Annular Lens (PAL) system [71], beneficial for eliminating the blind spot and expanding the field of view to 360° for real-world surrounding perception. To realize reliable segmentation in non-ideal and even adverse conditions, we aim to reduce the gap between conventional and omni-directional imaging-based street scene parsing, by exploiting the Panoramic Annular Semantic Segmentation (PASS) [8] dataset, as well as to bridge the daytime and nighttime image domains by using the ZJU driving dataset [72]. We will also widen the real-time performance comparison on newly released modules like Coral Dev Board and Nvidia Jetson Nano.

6. FUNDING INFORMATION

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 through the project "Research on Vision Sensor Technology Fusing Multidimensional Parameters" (111303-I21805) by Hangzhou SurImage Technology Co., Ltd and supported by Hangzhou KrVision Technology Co., Ltd (krvision.cn).

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. Krvision, "To Tackle the Challenges for the Visually Impaired," 2016, http://krvision.cn.
2. R. Girshick, "Fast R-CNN," in *International Conference on Computer Vision* (IEEE, 2015), pp. 1440-1448.
3. J. Long, E. Shelhamer and T. Darrel, "Fully Convolutional Networks for Semantic Segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2015), pp. 3431-3440.
4. K. Yang, L. M. Bergasa, E. Romera, R. Cheng, T. Chen and K. Wang, "Unifying terrain awareness through real-time semantic segmentation," in *IEEE Intelligent Vehicles Symposium* (IEEE, 2018), pp. 1033-1038.
5. K. Yang, K. Wang, L. M. Bergasa, E. Romera, W. Hu, D. Sun, J. Sun, T. Chen and E. López, "Unifying Terrain Awareness for the Visually Impaired through Real-Time Semantic Segmentation," Sensors **16**, 1-32 (2018).
6. M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth and B. Schiele, "The Cityscapes Dataset for

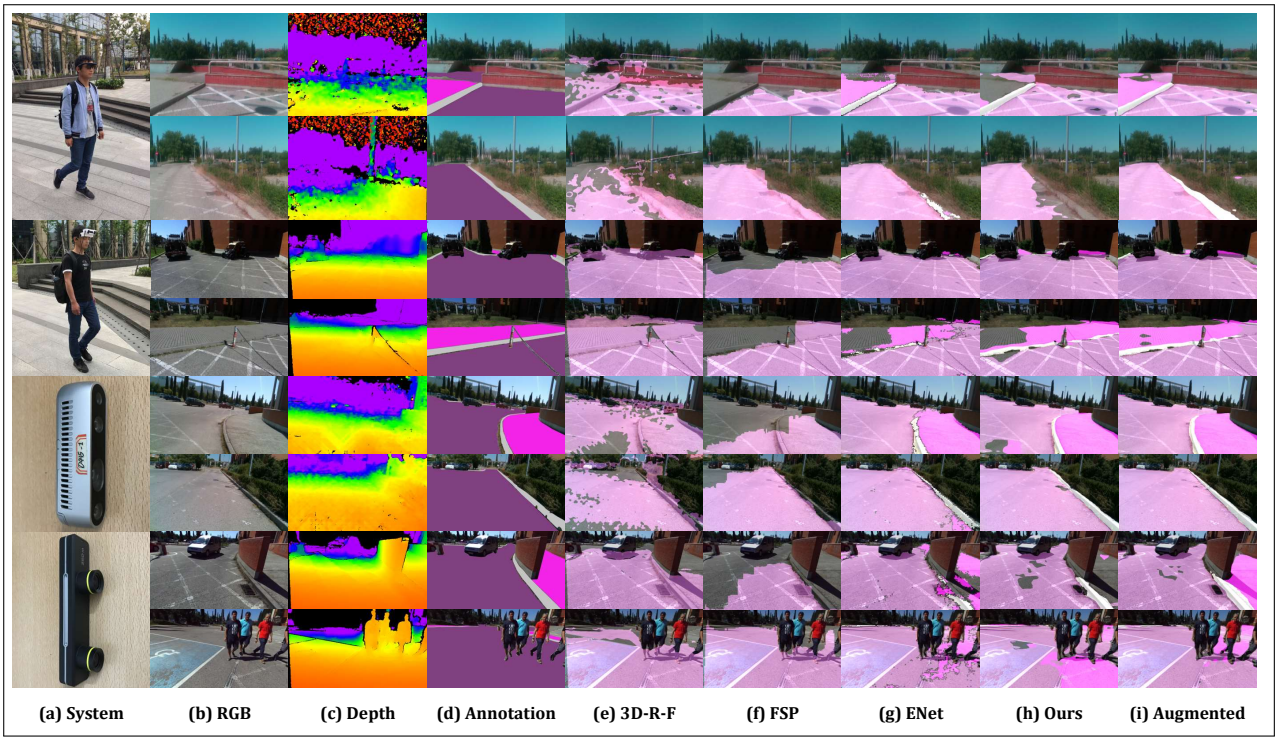

Fig. 7. Qualitative examples of the segmentation of real-world images produced by our approach (with/without data augmentations) compared with ground-truth annotation, 3D-RANSAC-F (3D-R-F) [15], FreeSpaceParse (FSP) [19] and ENet [28].

Semantic Urban Scene Understanding," in *IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2016), pp. 3213-3223.

7. G. Neuhold, T. Ollmann, S. R. Bulò and P. Kontschieder, "The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes," in *International Conference on Computer Vision* (IEEE, 2017), pp. 5000-5009.
8. K. Yang, "Panoramic Annular Semantic Segmentation" (github, 2019) [retrieved 20 March 2019], https://github.com/elnino9ykl/PASS.
9. G. Choe, S. H. Kim, S. Im, J. Y. Lee, S. G. Narasimhan and I. S. Kweon, "RANUS: RGB and NIR Urban Scene Dataset for Deep Scene Parsing," IEEE Robotics and Automation Letters **3**, 1808-1815 (2018).
10. F. Yu and V. Koltun, "Multi-Scale Context Aggregation by Dilated Convolutions," arXiv:1511.07122 (2015).
11. T. Pohlen, A. Hermans, M. Mathias and B. Leibe, "Full-Resolution Residual Networks for Semantic Segmentation in Street Scenes," in *IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2017), pp. 3309-3318.
12. L. C. Chen, Y. Zhu, G. Papandreous, F. Schroff and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," in *European Conference on Computer Vision* (2018), pp. 801-818.
13. Y. Zhuang, F. Yang, L. Tao, C. Ma, Z. Zhang, Y. Li, H. Jia, X. Xie and W. Gao, "Dense Relation Network: Learning Consistent and Context-Aware Representation for Semantic Image Segmentation," in *International Conference on Image Processing* (IEEE, 2018), pp. 3698-3702.
14. 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 *International Conference on Systems, Man, and Cybernetics* (IEEE, 2017), pp. 1399-1404.
15. 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**, 17467-17496 (2012).
16. K. Yang, K. Wang, W. Hu and J. Bai, "Expanding the Detection of Traversable Area with RealSense for the Visually Impaired," Sensors **16**, 1-20 (2016).
17. D. Pfeiffer and U. Franke, "Efficient Representation of Traffic Scenes by Means of Dynamic Stixels," in *Intelligent Vehicles Symposium* (IEEE, 2010), pp. 217-224.
18. M. Martinez, A. Roitberg, D. Koester, B. Stiefelhagen, and B. Schauerte, "Using Technology Developed for Autonomous Cars to Help Navigate Blind People," in *International Conference on Computer Vision Workshops* (IEEE, 2017), pp. 1424-1432.
19. 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 *International Conference on Robotics and Automation* (IEEE, 2017), pp. 6533-6540.
20. K. Yang, K. Wang, R. Cheng, W. Hu, X. Huang and J. Bai, "Detecting Traversable Area and Water Hazards for Visually Impaired with a pRGB-D Sensor," Sensors **17**, 1-20 (2016).
21. K. Yang, K. Wang, S. Lin, J. Bai, L. M. Bergasa and R. Arroyo, "Long-Range Traversability Awareness and Low-Lying Obstacle Negotiation with RealSense for the Visually Impaired," in *International Conference on Information Science and System* (ACM, 2018), pp. 137-141.
22. L. Deng, M. Yang, H. Li, B. Hu and C. Wang, "Restricted Deformable Convolution based Road Scene Semantic Segmentation Using Surround View Cameras," arXiv:1801.00708 (2018).
23. A. Sáez, L. M. Bergasa, E. Romera, E. López, R. Barea and R. Sanz, "CNN-based Fisheye Image Real-Time Semantic Segmentation," in *Intelligent Vehicles Symposium* (IEEE, 2018), pp. 1039-1044.
24. K. Yang, R. Cheng, L. M. Bergasa, E. Romera, K. Wang, N. Long, "Intersection perception through real-time semantic segmentation to assist navigation of visually impaired pedestrians," in *International Conference on Robotics and Biomimetics* (IEEE, 2018), pp. 1034-1039.
25. F. Schilling, X. Chen, J. Folkesson and P. Jensfelt, "Geometric and Visual Terrain Classification for Autonomous Mobile Navigation," in *International Conference on Intelligent Robots and Systems* (IEEE, 2017), pp. 2678-2684.
26. S. Ramos, S. Gehrig, P. Pinggera, U. Franke and C. Rother, "Detecting Unexpected Obstacles for Self-Driving Cars: Fusing Deep Learning and Geometric Modeling," in *IEEE Intelligent Vehicles Symposium* (IEEE, 2017), pp. 1025-1032.

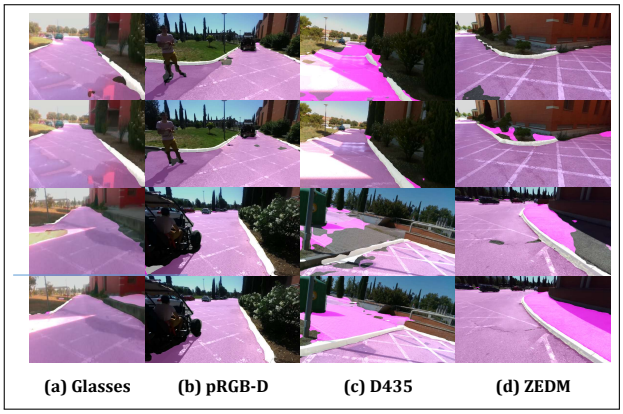

Fig. 8. Qualitative examples of the segmentation in unseen domains under challenging illumination and severe rotation conditions, which are produced by our approach with or without data augmentations. From left to right: images captured by (a) Smart Glasses, (b)pRGB-D Sensor, (c) RealSense D435 and (d) ZED Mini. The first and the third rows: semantic traversability perception results without data augmentations; the second and the fourth rows: segmentation outputs with data augmentations.

Fig. 9. Diverse qualitative examples of the segmentation of cross-country images (captured in Madrid, Spain and Hangzhou, China) produced by our semantic traversability perception approach with all augmentations. The color masks of road, sidewalks and curbs follow Vistas [7] style.

27. V. Badrinarayanan, A. Kendall and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence **39**, 2481-2495 (2017).
28. A. Paszke, A. Chaurasia, S. Kim and E. Culurciello, "ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation," arXiv:1606.02147 (2016).
29. A. Chaurasia and E. Culurciello, "LinkNet: Exploiting Encoder representations for Efficient Semantic Segmentation," in *IEEE Visual Communications and Image Processing* (IEEE, 2017), pp. 1-4.
30. 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**, 263-272 (2018).
31. H. Zhao, X. Qi, X. Shen, J. Shi and J. Jia, "ICnet for Real-Time Semantic Segmentation on High-Resolution Images," in *European Conference on Computer Vision* (2018), pp. 405-420.
32. M. Treml, J. Arjona-Medina, T. Unterthiner, R. Durgesh, F. Friedmann, P. Schuberth, A. Mayr, M. Heusel, M. Hofmarcher, M. Widrich and B. Nessler, "Speeding up Semantic Segmentation for Autonomous Driving," in Conference on Neural Information Processing System Workshop (2016), pp. 1-8.
33. W. Wang and Z. Pan, "DSNet for Real-Time Driving Scene Semantic Segmentation," arXiv:1812.07049 (2018).
34. R. P. Poudel, U. Bonde, S. Liwicki and C. Zach, "ContextNet: Exploring Context and Detail for Semantic Segmentation in Real-time," arXiv:1805.04554 (2018).
35. R. P. Poudel, S. Liwicki and R. Cipolla, "Fast-SCNN: Fast Semantic Segmentation Network," arXiv:1902.04502 (2019).
36. Q. Ning, J. Zhu and C. Chen, "Very Fast Semantic Image Segmentation Using Hierarchical Dilation and Feature Refining," Cognitive Computation **10**, 62-72 (2018).
37. L. Tang, X. Ding, H. Yin, Y. Wang and R. Xiong, "From one to many: Unsupervised traversable area segmentation in off-road environment," in *International Conference on Robotics and Biomimetics* (IEEE, 2017), pp. 787-792.
38. Y. H. Tsai, W. C. Hung, S. Schulter, K. Sohn, M. H. Yang and M. Chandraker, "Learning to Adapt Structured Output Space for Semantic Segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2018), pp. 7472-7481.
39. Y. Zhang, H. Chen, Y. He, M. Ye, X. Cai and D. Zhang, "Road segmentation for all-day outdoor robot navigation," Neurocomputing **314**, 316-325 (2018).
40. S. Mehta, H. Hajishirzi and L. Shapiro, "Identifying Most Walkable Direction for Navigation in an Outdoor Environment," arXiv:1711.08040 (2017).
41. D. K. Kim, D. Maturana, M. Uenoyama and S. Scherer, "Season-Invariant Semantic Segmentation with a Deep Multimodal Network," in Field and Service Robotics, 255-270 (2018).
42. A. Valada, J. Vertens, A. Dhall and W. Burgard, "Adapnet: Adaptive Semantic Segmentation in Adverse Environmental Conditions," in *International Conference on Robotics and Automation* (IEEE, 2017), pp. 4644-4651.
43. K. Yang, L. M. Bergasa, E. Romera, X. Huang, and K. Wang, "Predicting polarization beyond semantics for wearable robotics," in *International Conference on Humanoid Robots* (IEEE, 2018), pp. 96-103.
44. N. Alshammari, S. Akcay and T. P. Breckon, "On the Impact of Illumination Invariant Image Pre-transformation on Contemoporary Automotive Semantic Scene Understanding," in *IEEE Intelligent Vehicles Symposium* (IEEE, 2018), pp. 1027-1032.
45. C. Sakaridis, D. Dai, S. Hecker and L. Van Gool, "Model Adaptation with Synthetic and Real Data for Semantic Dense Foggy Scene Understanding," in *European Conference on Computer Vision* (2018), pp. 707-724.
46. C. Sakaridis, D. Dai and L. Van Gool, "Semantic Nighttime Image Segmentation with Synthetic Stylized Data, Gradual Adaptation and Uncertainty-Aware Evaluation," arXiv:1901.05946.
47. A. Arnab, O. Miksik and P. H. Torr, "On the Robustness of Semantic Segmentation Models to Adversarial Attacks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 888-897.
48. G. L. Oliveira, C. Bollen, W. Burgard and T. Brox, "Efficient and Robust Deep Networks for Semantic Segmentation," The International Journal of Robotics Research **37**, 472-491 (2018).
49. O. Zendel, K. Honauer, M. Murschitz, D. Steininger and G. F. Dominguez, "WildDash-Creating Hazard-Aware Benchmarks," in *European Conference on Computer Vision* (2018), pp. 402-416.
50. M. Larsson, E. Stenborg, L. Hammarstrand, T. Sattler, M. Pollefeys and F. Kahl, "A Cross-Season Correspondence Dataset for Robust Semantic Segmentation," arXiv:1903.06916 (2019).
51. L. E. Ortiz, E. V. Cabrera and L. M. Gonçalves, "Depth Data Error Modeling of the ZED 3D Vision Sensor from Stereolabs," ELCVIA: electronic letters on computer vision and image analysis **17**, 0001-15 (2018).
52. L. Keselman, J. Iselin Woodfill, A. Grunnet-Jepsen and A. Bhowmik, "Intel RealSense Stereoscopic Depth Cameras," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops* (IEEE, 2018),

pp. 1-10.
53. S. Giancola, M. Valenti and R. Sala, “Metrological Qualification of the Intel D400™ Active Stereoscopy Cameras,” in *A Survey on 3D Cameras: Metrological Comparison of Time-of-Flight, Structured-Light and Active Stereoscopy Technologies* (Springer, 2018), pp. 71-85.
54. Stereolabs, “Meet ZED Mini: the world’s first camera for mixed-realit,” 2018, https://www.stereolabs.com/zed-mini.
55. 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**, 2809-2819 (2018).
56. 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 **9**, 743-755 (2017).
57. K. Yang, “Sequential/Hierarchical ERF-PSPNet” (github, 2018) [retrieved 24 December 2018], https://github.com/elnino9ykl/ERF-PSPNet.
58. K. Xiang, “ERF-PSPNET implemented by tensorflow” (github, 2018) [retrieved 24 December 2018], https://github.com/Katexiang/ERF-PSPNET.
59. D. Sun, “ERF-PSPNet on TX2” (github, 2018) [retrieved 24 December 2018], https://github.com/dongmingsun/tx2-erfpsp.
60. K. He, X. Zhang, S. Ren and J. Sun, “Deep Residual Learning for Image Recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2016), pp. 770-778.
61. H. Zhao, J. Shi, X. Qi, X. Wang and J. Jia, “Pyramid Scene Parsing Network,” in *IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2017), pp. 6230-6239.
62. E. Romera, L. M. Bergasa, J. M. Alvarez and M. Trivedi, “Train Here, Deploy There: Robust Segmentation in Unseen Domains,” in *IEEE Intelligent Vehicles Symposium* (IEEE, 2018), pp. 1823-1833.
63. K. Wang, “RGB-D-SS Dataset,” (2018) [retrieved 24 December 2018], http://wangkaiwei.org/downloadeg.html.
64. D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” arXiv:1412.6980 (2014).
65. T. Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, “Focal Loss for Dense Object Detection,” in *International Conference on Computer Vision* (2017, IEEE), pp. 2990-3007.
66. X. Y. Zhou, C. Riga, S. L. Lee and G. Z. Yang, “Towards Automatic 3D Shape Instantiation for Deployed Stent Grafts: 2D Multiple-class and Class-imbalance Marker Segmentation with Equally-weighted Focal U-Net,” in *International Conference on Intelligent Robots and Systems* (IEEE, 2018), pp. 1261-1267.
67. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg and F. Li, “Imagenet Large Scale Visual Recognition Challenge,” International Journal of Computer Vision **115**, 211-252 (2015).
68. K. He, J. Sun and X. Tang, “Guided Image Filtering,” in *European Conference on Computer Vision* (2010), pp. 1-14.
69. L. He, G. Wang and Z. Hu, “Learning Depth from Single Images with Deep Neural Network Embedding Focal Length,” IEEE Transactions on Image Processing **27**, 4676-4689 (2018).
70. B. Chen, C. Gong, and J. Yang, “Importance-Aware Semantic Segmentation for Autonomous Vehicles,” IEEE Transactions on Intelligent Transportation Systems **20**, 137-148 (2019).
71. Y. Luo, X. Huang, J. Bai and R. Liang, “Compact polarization-based dual-view panoramic lens,” Applied Optics **56**, 6283-6287 (2017).
72. K. Yang, “ZJU Day and Night Driving Dataset” (github, 2019) [retrieved 20 March 2019], https://github.com/elnino9ykl/ZJU-Dataset.