Predicting polarization beyond semantics for wearable robotics

Kailun Yang[1], Luis M. Bergasa[2], Eduardo Romera[2], Xiao Huang[3] and Kaiwei Wang[1]

Abstract— **Semantic perception is a key enabler in robotics, which supposes a very resourceful and efficient manner of applying vision information for upper-level navigation and manipulation tasks. Given the challenges on specular semantics such as water hazards, transparent glasses and metallic surfaces, polarization imaging has been explored to complement the RGB-based pixel-wise semantic segmentation because it reflects surface characteristics and provides additional attributes. However, polarimetric measurements generally entail prohibitively expensive cameras and highly accurate calibrations. Inspired by the representation power of Convolutional Neural Networks (CNNs), we propose to predict polarization information from monocular RGB images, precisely per-pixel polarization difference. The core of our approach is a cluster of efficient deep architectures building on factorized convolutions, hierarchical dilations and pyramid representations, aimed to produce both semantic and polarimetric estimations in real time. Comprehensive experiments demonstrate the qualified accuracy on a wearable exoskeleton humanoid robot.**

I. INTRODUCTION

Semantic classification [1] is a fundamental topic in robotic perception systems. The segmentation process, posed as per-pixel prediction to divide observed scenes into semantic regions, has become the key enabler to unify monocular detectors for autonomously driving vehicle [2], assistive wearable device [3], and humanoid robot [4] performing locomotion, navigation or manipulation tasks. Notably, two essential challenges that still need to be addressed are: 1) to efficiently achieve accurate semantic segmentation in real time, and 2) to effectively deal with critical semantic surfaces in real world. An example could be the underlying water hazards shown in Fig. 1, which are very dangerous for robots but they tend to be classified as general road markings by most of the current perception systems.

Polarization and its imaging extend the information dimension to be used for target detection [5], with the complementary characteristics coded implicitly in the polarization state of light when reflected from different specular surfaces such as water puddles [6], transparent glasses [7] and metallic materials [8]. However, polarimetric measurements generally require expensive micro-polarizer cameras, multi-view observations and precise calibrations. More specifically, a micro-polarizer camera has been made by 4D Technology [9] thanks to the development of polarization-sensitive focal plane arrays. It costs more than 20000USD and depends on the careful calibration process, which is crucial to its polarimetric measurement performance [10]. While such micro-polarizer camera has been applied for microscopes [11] and interferometers [12] due to the capability to sense under four polarization angles with a single shot (see Fig. 1(a)), it remains comparatively cost-prohibitive for robotics community, thus only seen in limited work [7] [13].

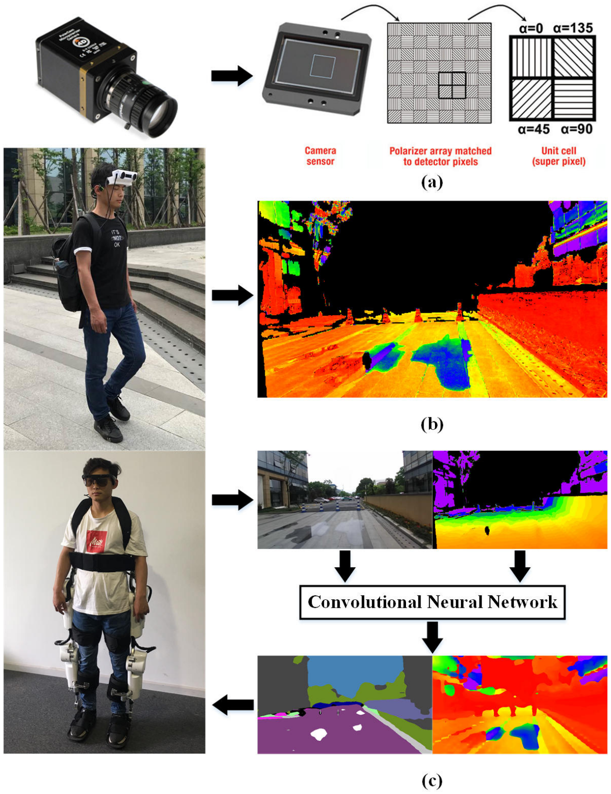

α=0
α=135
α=45
α=90
Camera sensor
Polarizer array matched to detector pixels
Unit cell (super pixel)
(a)
(b)
Convolutional Neural Network
(c)

Fig. 1. Three approaches of polarization perception: (a) Micro-grid array for polarization camera [9], (b) Polarization difference imaging with a polarized stereo camera [6], (c) Proposed approach for wearable robotics: polarization prediction beyond semantics from RGB images (alternatively together with depth images) using a convolutional neural network.

As a workaround to this issue, the degree of linear polarization in polarization difference imaging [14] has been widely employed as a basis of polarization analysis for perception and reconstruction of real-world specular scenes [6] [7] [8]. However, polarization difference is captured through multi-view or time-division imaging, which is tightly intertwined with the precise rotation of polarizers attached on different cameras. These factors all limit the flexibility and applicability, especially for wearable robotics.

Nowadays, Convolutional Neural Networks (CNNs) learn and discriminate between different features directly from the

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

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

[3] Xiao Huang is with College of Optical Sciences, University of Arizona, Tucson, AZ, USA xhuang@optics.arizona.edu

input data using a deeper abstraction of representation layers. Promisingly, recent advances in deep learning have achieved break-through results in semantic segmentation [15], depth estimation [16] and multi-focus image fusion [17]. Taking inspiration from such representation power of CNNs, we propose to predict polarization information from monocular RGB images in an end-to-end, fully convolutional manner. To the best of our knowledge, this is the first time dense pixel-wise polarimetric measurements are inferred from light intensity-based imagery. Our system can be used for military and industrial purposes to enhance user mobility through augmented/virtual reality interactions, e.g., highlighting the critical hazards on the navigational path, or the whole path in scenarios under poor illumination. In this way, exoskeleton humanoid robots can also assist disabled people in their rehabilitation training, or visually impaired users in everyday self-navigation by using acoustic feedback.

The primary contribution of the paper is our polarization prediction proposal. In addition, novel technical contributions and results reside in the following aspects:

- A cluster of real-time architecture instances building on factorized convolutions, hierarchical dilations and pyramid representations, which can be used for semantic/polarimetric prediction. PyTorch codes corresponding to these architectures will be open-sourced at *github.com/elnino9ykl/ERF-PSPNet*.
- A suit of real-world wearable assistive navigation system (see Fig. 1(c)) including an exoskeleton robot and a pair of smart glasses made available at *krvision.cn*.
- A comprehensive set of experiments on a large-scale scene parsing dataset and an ego-centric campus dataset, which can be accessed at *wangkaiwei.org/downloadeg.html*.

The remainder of this paper is structured as follows. Section II reviews the literature mainly related to semantic segmentation and polarization imaging for wearable robotics. In Section III, the perception framework is elaborated. In Section IV, the approach is evaluated and discussed as for real-time/real-world performance of semantic/polarimetric prediction. Section V draws the conclusions and gives an outlook to future research.

II. RELATED WORK

Semantic classification has been fueled by the progressively emerging deep learning pipelines and architectures. Among the literature, a vital part of networks are predominantly based on FCNs [15], which pioneered the era of end-to-end segmentation. SegNet [18] revolutionized efficient semantic segmentation by symmetrizing the encoder-decoder design. ENet [19] implemented real-time semantic segmentation by employing a larger encoder as good feature extractors and the decoder is only responsible for fine-tuning details. This simplified structure sacrifices a good deal of accuracy in order to remain efficient. In our previous work, ERFNet [2] pursued the aim of maximizing the trade-off between accuracy/efficiency in case CNN-based segmentation needs to be applied on resource-constrained edge devices. In a similar spirit, LinkNet [20] attempted to achieve accurate instance-level prediction without compromising processing time by linking the encoder and the corresponding decoder. Recently, a very fast segmentation approach introduced cognitive multi-scale hierarchical dilation, while exploiting intermediate supervision to refine coarse features [21]. Although these architectures claimed to have less performance drop along with the impressing speedup, most of them are designed for autonomous vehicles. In the state of the art, real-time architectures tailored for assistive wearable robotics are scarce, which is a time-critical, context-critical and safety-critical domain. More importantly, RGB-based segmentation degrades on specular surfaces, which are commonplace during everyday navigation.

Polarization imaging is convincingly appealing to enhance semantic segmentation although human vision has not evolved to exploit polarimetric information, such that the complementarity has not been thoroughly confirmed. Along with the advent of compact micro-polarizer cameras that integrate micro-grid polarization filter arrays on the focal plane to provide four orientations of linear polarization (see Fig. 1(a)), stereo and polarization cues were incorporated in a unified formulation [7]. Although related, it relies on the synthetically generated data to model specular parts of indoor scenes. In real-world setting, we previously presented a pRGB-D framework with a polarized stereo camera to unify the segmentation of traversable areas and water hazards for the visually impaired [6]. Although the coverage of assistance has been extended to vulnerable pedestrians, polarized stereo cameras were also used in advanced driver assistance systems [22] [23]. Within the intelligent vehicle context, polarization information was utilized as features by leveraging classical methods, in order to tackle the problem of outdoor scene segmentation on objects with strong reflection or poor illumination [5]. While describing imperceptible light properties, the policies in these multi-modality frameworks remain largely handcrafted.

Wearable robotics is a research field with vast literature, but it is seldom bridged with data-driven semantic perception. A team of researchers proposed the semantic paintbrush [24], which is an augmented reality system based on RGB-IR stereo setup and optical see-through headset, along with a laser pointer allowing the user to draw directly onto its 3D reconstruction. Unlike typical assistive systems, it places the user in the loop to exhaustively segment semantics of interest. In the domain of cycling, HindSight [25] uses a deep neural network to locate and attribute semantic information to objects surrounding a cyclist through a head-worn panoramic camera. While inspiring, this work focused on redirecting attentions to sonified objects by providing only sparse bounding-box semantic predictions and hence cannot be directly used for upper-level tasks. Similar bounding-box interpretation was addressed when ultrasonic sensors and computer vision joined forces [26]. Based on a wearable smartphone waist belt, it semantically interpreted the detected obstructions based on their degree of danger, with the aim to facilitate autonomous navigation of blind people

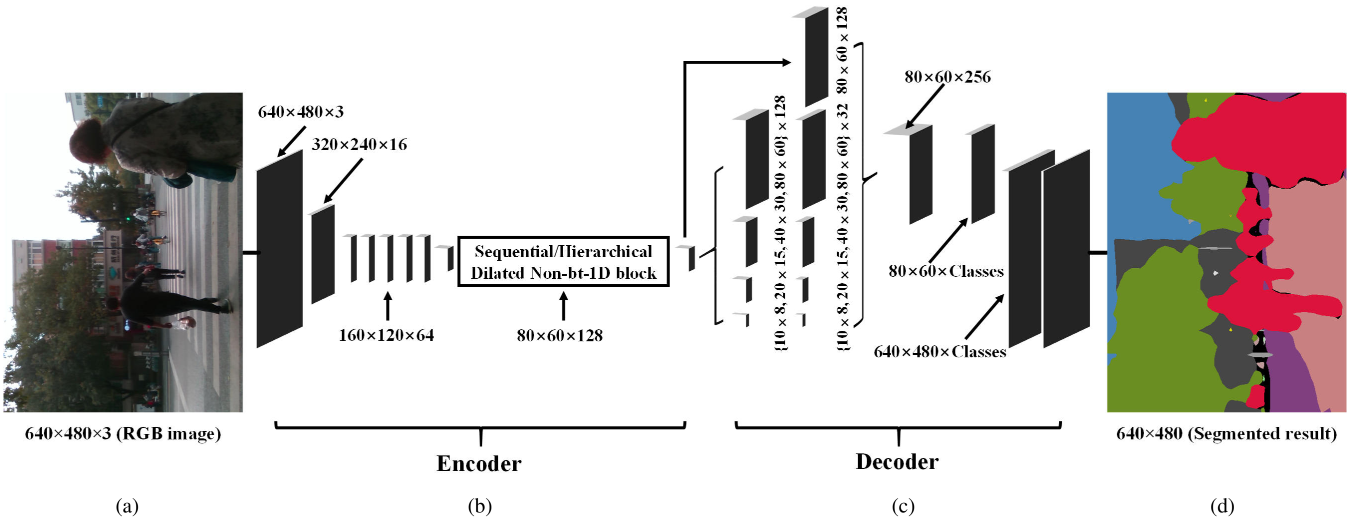


Fig. 2. The proposed architecture. From left to right: (a) Input, (b) Encoder, (c) Decoder, (d) Prediction.

in highly dynamic urban scenes, which is limited to only four categories. In contrast, we target at dense and diverse predictions, which suppose a very rich source of processed information, including per-pixel polarimetric and semantic estimations attaining coverage of various traffic elements.

III. APPROACH

A. System and architecture overview

In this research, the main motivation is to design a prototype which should be wearable without hurting the self-esteem of users with some disability. With this target in mind, we follow the trend of using head-mounted glasses [27] to acquire environmental information and interact with the user. As worn by the user in Fig. 1(c), the pair of 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 (alternatively together with depth images) are fed to the network for semantic/polarimetric estimation. Based on the per-pixel prediction, augmented reality or assistive awareness can be rendered, e.g., acoustic feedback through the bone conducting earphones. In addition, other sensors and accelerometers built in the exoskeleton robot collect the motion and muscle signals with the eventual goal to customize walking modes for the users. In this work, we focus on visual perception of the exoskeleton humanoid robot system.

Up until very recently, the applicability of per-pixel semantic scene parsing on embedded devices is limited by the speed. However, efficient end-to-end segmentation has been an intriguing research topic with a growing community focused on designing networks [2] [18] [19] that could parse the entire scene in near real time. These advances have made possible the utilization of pixel-level inference in time-critical applications like semantic/polarimetric estimation within wearable robotics context. Towards this objective, our architecture is designed according to the encoder-decoder architectures like SegNet [18], ENet [19] and our previous ERFNet [2]. In FCN-like architectures [15], feature maps from different layers need to be fused to generate a fine-grained output. As expanded in Fig. 2, our approach uses a more sequential architecture based on an encoder producing down-sampled feature maps, and a subsequent decoder that up-samples feature maps to match input resolution. Table I also gives a detailed description of the proposed architecture building on factorized convolutions, hierarchical dilations and pyramid representations, which will be elaborated in the following subsections.

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

	Layer	Type	Out-F	Out-Res
ENCODER	0	Original RGB/RGB-D image	3/4	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

B. Factorized convolution

Generally speaking, the residual layer adopted in state-of-the-art networks has two instances, the bottleneck version and the non-bottleneck design. Based on 1D factorizations of the convolutional kernels, "Non-bottleneck-1D" (Non-bt-1D) was redesigned in our previous work [2] [3] to strike a rational balance between the efficiency of bottleneck and the learning capacity of non-bottleneck. Precisely, the spatial factorization into separable asymmetric convolutions is leveraged to improve the computational efficiency of the 3×3 convolutions in the residual modules. Upon the addition of down-sampler block inspired by ENet [19] that concatenates the parallel outputs of a single 3×3 convolution with stride 2 and a max-pooling module, our encoder enables an efficient

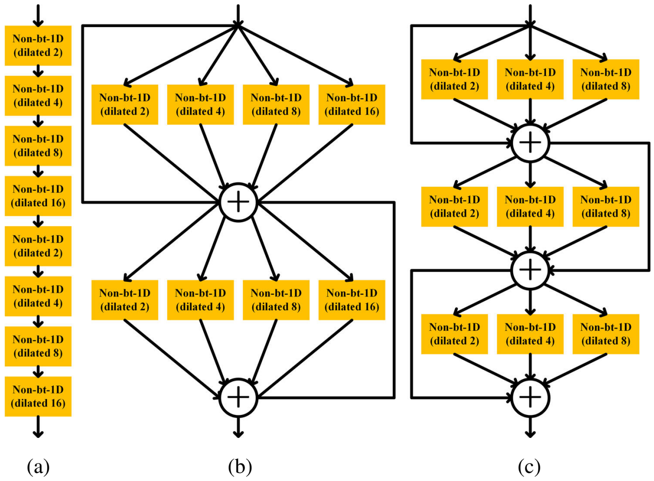


Fig. 3. Sequential and hierarchical architectures of dilated Non-bottleneck-1D (Non-bt-1D) layers. From left to right: (a) Sequential architecture, (b) 4×2 hierarchical architecture, (c) 3×3 hierarchical architecture.

use of residual layers to extract feature maps and further achieve semantic/polarimetric estimation in real time.

C. Hierarchical dilation

Starting from the observation that increased number of layers helps to learn more complex and abstract features, which leads to increased 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 Fig. 3. Compared with conventional schemes [21] [28], 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) in diverse sizes, 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 leads to an implicit deep supervision, so 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 [2] [3].

D. Pyramid representation

Another key modification lies in the decoder with specific insights about contextual information and its cardinal significance for the prediction of semantic/polarimetric information. To detail this, two common issues are worthwhile to remark for context-critical wearable robotics. Firstly, if the network mis-predicts curbs on crosswalks, the wearer would be left vulnerable in the dynamic environments given such feedback. The common knowledge should be learned by the data-driven approach that curbs are seldom over crosswalks, and the specular water hazards that exhibit higher polarization difference are generally encompassed by traversable areas. Secondly, when navigating the sidewalks or crossing the roads, the scene elements such as vehicles with metallic surfaces, hazardous curbs and water puddles will exhibit arbitrary sizes from the sensor perspective. Wearable robotic system should pay much attention to different sub-regions that contain inconspicuous-category stuff.

Learning more relationship between scene categories by exploiting more context represents a promising approach to mitigate these risks. In this reconstruction, the decoder architecture follows the pyramid pooling module as introduced by PSPNet [29]. This module is leveraged to harvest different sub-region representations, followed by up-sampling and concatenation layers to form the final feature representations. In this manner, local and global context information are carried from the pooled representations at different locations. By fusing features under a group of different pyramid levels, the output of different levels in this pyramid pooling module contains the feature map from the encoder with varied sizes. As shown in Fig. 2(c), the weight of global feature from the encoder is maintained, with 1×1 point-wise convolution layer appended after each pyramid level to reduce the dimension of context representation. Subsequently, the low-dimension feature maps are directly up-sampled to obtain the same size features as the original feature map through bilinear interpolation. Overall, Fig. 2 contains a depiction of the feature maps generated by each of the block in our architecture, from the RGB input to the per-pixel class probabilities and final prediction.

IV. EXPERIMENTS

Experiment setup. The experiments are performed in public spaces around Holley Metering Campus in Hangzhou. A real-world pRGB-D dataset is captured using the polarized stereo camera [6]. It is a commercial stereo camera that has been the driving force behind the success of many robotic vision algorithms, owing to its outstanding ability to perform large-scale depth perception at up to 20m. We retrofit it by attaching horizontal and vertical polarization filters on the left and right camera respectively. In this fashion, we have 9736 images with per-pixel ground-truth polarization difference, which is obtained by warping the right image to the left image to produce point correspondence using the disparity that can be directly generated from the dense depth image in the first place. Subsequently, to construct a representation of polarimetric information, each polarized stereo pair is exploited to calculate a pixel-wise brightness difference image that indicates the degree of polarization resulting from reflection. The dataset is randomly separated into a training subset with 8758 images and a testing subset with 978 images.

As far as navigational semantic segmentation is concerned, the challenging Mapillary Vistas dataset [30] is chosen as it consists of many traversability-related classes, featuring a high variability in capturing viewpoints and spanning a broad range of outdoor scenes on different pathways or sidewalks, which corresponds to the usage scenario of the wearable humanoid robot. In total, we have 18000 images for training and 2000 images for validation with pixel-exact annotations.

The metrics reported in this paper correspond to Intersection-over-Union (IoU) that is prevailing in semantic segmentation challenges, and following metrics originally used in the field of depth estimation [16]:

- RMSE : root mean squared error.
- REL : mean absolute relative error.
- δ_i : percentage of predicted pixels where the relative error is within a threshold. Formally put,

$$\delta_i \quad = \quad \frac{\mathbf{card}\Big(\Big\{\hat{y}_i : \max\Big\{\frac{\hat{y}_i}{y_i}, \frac{y_i}{\hat{y}_i}\Big\}\Big\} < 1.25^i\Big)}{\mathbf{card}(\{y_i\})}, \tag{1}$$

where y_i and $\hat{y}_i$ are respectively the ground-truth value and the prediction value, and **card** is the cardinality of a set. In this regard, a higher δ_i indicates better prediction.

Training setup. To provide the assistive awareness regarding the semantics of interest 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. 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. Additionally, random rotation by sampling distributions from the ranges [-20^{o}, 20^{o}], color jittering from the ranges [-0.2, 0.2] for hue, [0.8, 1.2] for brightness, saturation and contrast are also applied. Our model is trained using Adam optimization, initiated with a learning rate of $5{\times}10^{-5}$ that decreases exponentially across epochs. First the encoder is trained by mapping an input to a down-sampled label, then the corresponding decoder is appended to the trained encoder to perform up-sampling and train the overall network followed by a pixel-wise classifier. Following the weight determining scheme in [19], the training of the full network reaches convergence when focal loss [31] is adopted as the criterion:

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

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 (when $\mathbf{P}_{(i,j,n)} = 0.1$, $(1-\mathbf{P}_{(i,j,n)})^2$=0.81). In this way, the focal loss concentrates the training on wrongly-segmented pixels or hard pixels.

Comparatively, for polarization prediction, the network is trained by mapping the RGB or RGB-D input to 255 levels of polarization difference using cross-entropy loss with uniform weights for all valid levels. Data augmentations used in this training only involve random flipping, cropping and scaling with same probability setting as semantic segmentation, so that the key idea of inferring polarization information from light intensity-based imagery is validated in a purely experimental manner.

Real-time performance. As displayed in Table II, the frame rates of our sequential/hierarchical ERF-PSPNets are tested and compared with the state-of-the-art networks for real-time semantic segmentation including ENet [19] and LinkNet [20]. At 320×240, a resolution that is enough to recognize any urban scene accurately and create augmented visual/acoustic reality for the wearable humanoid robot, 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. However, for navigation assistance, 320×240 is arguably the optimum resolution of the three resolutions, since pixel-exact features are less desired by the user, but necessitate higher input resolution that incurs longer processing latency. Still, the mean IoU values of our models tested on Mapillary dataset [30] 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.

TABLE II
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 [19]	66.2FPS	57.5FPS	**41.8FPS**	33.6%
LinkNet [20]	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 while consuming less than 10 Watts at full load, and our models achieve more than 22.0FPS at 320×240. Evidently, the hierarchical architectures are both faster than the sequential architecture while only causing a minor drop in segmentation performance.

Segmentation accuracy. Table III 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-PSPNets exceed the existing architectures that are also designed for real-time applications by a wide margin. 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 traversablity and traffic safety perception are both higher.

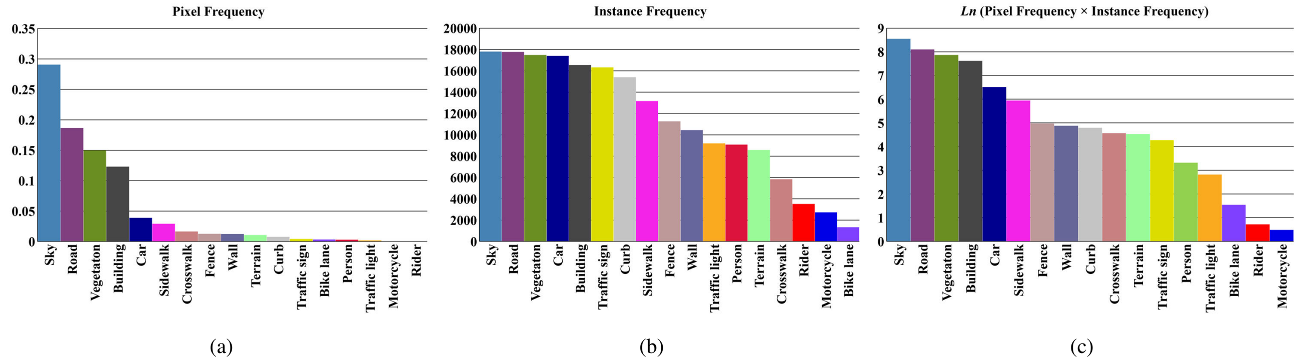
When comparing the hierarchical ERF-PSPNets with the sequential version, although the mean IoU value of all classes used for training is lower, they offer some benefits. Firstly, in spite of being a possible subjective measure, the mean IoU values of 17 main navigational classes are higher than that achieved with the sequential design. Secondly, on some

TABLE III

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 [19]	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%	94.4%	81.1%	52.9%	51.0%	33.6%	45.6%
LinkNet [20]	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%**



(a) (b) (c)

Fig. 4. Class frequency of the Mapillary dataset. From left to right: (a) Pixel frequency (proportion of labeled pixels), (b) Instance frequency (number of images with at least one labeled instance), (c) Natural logarithm of the product of pixel frequency and instance frequency.

semantic classes of interest, hierarchical designs achieve best accuracy. 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 Fig. 4(b)(c), *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 version reaches up to 6.8%. It suggests that the proposed HD-1D block is promising to boost the segmentation performance on less frequent classes. This is a very positive aspect because it is always harder to achieve good performance on less frequent classes. For example, ENet fails to classify *Motorcycle* well as its accuracy is only 0.1%. Such problem of ENet was also reported in [21]. On application side, hierarchical versions of ERF-PSPNet could enhance the safety of humanoid robots in streets/lanes with many motorcycles/bicycles that seriously influence the traffic flow, especially within metropolitan areas of China cities that have implemented public bicycle-sharing programs as a strategy to promote low-carbon transportation. Noticeably, the performance gap is more related with instance frequency rather than pixel frequency as illustrated in Fig. 4.

It is worthwhile to mention that we have also pre-trained the sequential version of ERF-PSPNet on ImageNet, eventually the mean IoU value reaches 48.8%, which is marginally higher than that achieved with the "from scratch" strategy (48.4%). This result reveals that although the transferability of features of pre-training on a large dataset is advantageous, our models can also reach good accuracy when trained on a single dataset without the need of pre-training that adds training complexity and may suppose commercial limitations.

Polarization prediction. Taking an essential step further than semantic segmentation, the produced polarimetric estimations are evaluated on the real-world pRGB-D dataset captured with the polarized stereo sensor [6]. The reported results in Table IV involve error metrics and accuracy metrics. The error metrics are more related with spatial details/representations, while the accuracy metrics expect the model to deliver correct distribution of polarimetric information structurally and contextually. Accordingly, the sequential version of ERF-PSPNet achieves best results on accuracy metrics, since the directly stacked residual layers suppose larger depth and better capability to gather abstract global context information. However, the 4×2 hierarchical version yields better performance on error metrics, because the bypass connection allows the network to use less levels of structure to capture features from diverse sizes of FoV, which help to infer correct spatial details. In contrast, the 3×3 hierarchical design is under-performing, which implies that the removed Non-bt-1D with 16 dilation rate (see Fig. 3(c)) still matters in polarization prediction.

TABLE IV

POLARIZATION PREDICTION ANALYSIS.

Network/Input	**Error metrics (lower, better)**		**Accuracy metrics (higher, better)**		
	RMSE	**REL**	δ_1	δ_2	δ_3
ENet [19]	13.87	6.37	46.82%	62.73%	68.85%
LinkNet [20]	13.30	6.42	45.47%	61.62%	67.93%
Seq	13.35	6.14	**48.83%**	**63.85%**	**69.41%**
4×2 Hier	**13.17**	**6.09**	48.59%	63.32%	68.85%
3×3 Hier	13.46	6.23	47.71%	62.94%	68.68%
RGB	13.35	6.14	48.83%	63.85%	69.41%
RGB-D	**13.22**	**6.08**	48.53%	63.75%	69.56%
HSV	13.56	6.39	47.24%	62.54%	68.36%
HSV-D	13.46	6.24	48.02%	63.65%	69.58%
V	13.29	6.18	48.09%	63.25%	68.82%
V-D	13.24	6.10	**48.94%**	**64.00%**	**69.64%**

Loosely speaking, polarization difference imaging is more related to brightness, so it would be straightforward to predict it from HSV or only V channel using an alternative input representation. Based on this intuition, we perform the experiments to train and test by feeding/inputting the network in different representations using the sequential ERF-PSPNet. According to the results visualized in Table IV, we arrive three findings. First, the V-D inputting achieves best results on accuracy metrics which fits our expectations. Second, on error metrics the RGB-D inputting achieves better results which may be ascribed to the additional information favorable to polarization classification, especially on specular scene parts. Third, it is consistent that RGB-D inputting outperforms RGB, while HSV-D outperforms HSV and V-D outperforms V-based prediction. This means depth channel has played a role because polarization is related with not only surface characteristics but also range information. For illustration, we have derived that the polarization difference cue becomes strong at distances above a minimum range in a previous pRGB-D perception framework [6].

Qualitative analysis. Fig. 5 exhibits the montage of pixel-wise results generated by our approach, LinkNet and ENet. Not surprisingly, our approach not only yields longer and more consistent semantic segmentation which will definitely benefit the traversability-related awareness of different pathways and hazardous curbs, but also retains outstanding ability to perceive traffic risks with regard to vehicles and pedestrians. On specular surfaces, such as water puddles, these models all suffer from a degradation of performance when facing such common yet unseen scenarios, especially the LinkNet, although it achieves higher accuracy than ENet. The main insight gained from our experiment 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 and annotation diversity as well as information dimensionality. More critically, neglecting such real-world conditions impairs the overall performance of semantic segmentation and incurs a bias of the appearance of scene elements to be analyzed.

The predicted per-pixel polarization difference image is promising for conquering the problems on specular regions in the real world. As depicted in Fig. 5, our approach produces smooth polarimetric estimations of water hazards/wet areas on the navigational path, transparent glasses on the parked cars, and windows of the modern buildings, etc. Strikingly, the network even predicts dense polarization information that is unable to be captured by the polarized stereo sensor, e.g., the water areas in the first row and the windows in the third row, which visually evidences the generalization capability. When comparing the qualitative performance between RGB/RGB-D inputting, it is true that the fed RGB-D information helps to perceive more details, but it also brings noises, especially at far-away ranges when depth information become sparse and less reliable from the perspective of our humanoid robot. Still, it is encouraging that robotic perception based on pixel-wise semantic segmentation can be enhanced to a great extent through polarization prediction.

V. CONCLUSIONS

Fueled by deep learning, semantic segmentation has become a de-facto standard in robotics. To complement real-world semantic classification on specular surfaces, we propose to estimate per-pixel polarization information, which can be used to prevent from potential hazards for wearable humanoid robots, especially safety-critical ones fulfilling navigational tasks, e.g., our assistive exoskeleton with augmented reality glasses. Traditionally, polarization information is independent with light intensity-based information such as RGB and depth data. Our work bridges the optical polarization and intensity information through real-time end-to-end prediction using a cluster of customized efficient networks. Based on our proposal, it is highly portable to apply semantic and polarimetric estimations on embedded devices by using only a single RGB camera, without resorting to expensive micro-polarizer sensors, or multiple cameras attached with rotatable polarization filters.

We are aware that there is much room for improvement in this research line, e.g., the indoor robustness. It is planned to collect larger ground-truth dataset with high-end polarized sensor that can produce smooth direct polarimetric measurements, and make our approach generalize beyond polarization difference imaging in a broad variety of scenarios. Another promising direction is to experiment with different loss functions to provide meaningful training supervision, and to include depth-wise separable convolution as an additional shallow branch for spatial detail fine-tuning.

ACKNOWLEDGMENT

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

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

REFERENCES

[1] A. Nüchter and J. Hertzberg, Towards semantic maps for mobile robots, *Robotics and Autonomous Systems*, 56(11), 2008, pp. 915-926.

[2] E. Romera, J.M. Alvarez, L.M. Bergasa and R. Arroyo, Erfnet: Efficient Residual Factorized ConvNet for Real-Time Semantic Segmentation, *IEEE Transactions on Intelligent Transportation Systems*, 19(1), 2018, pp. 263-272.

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

[4] M. Brandao, Y.M. Shiguematsu, K. Hashimoto and A. Takanishi, Material recognition CNNs and hierarchical planning for biped robot locomotion on slippery terrain, In *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*, 2016, pp. 81-88.

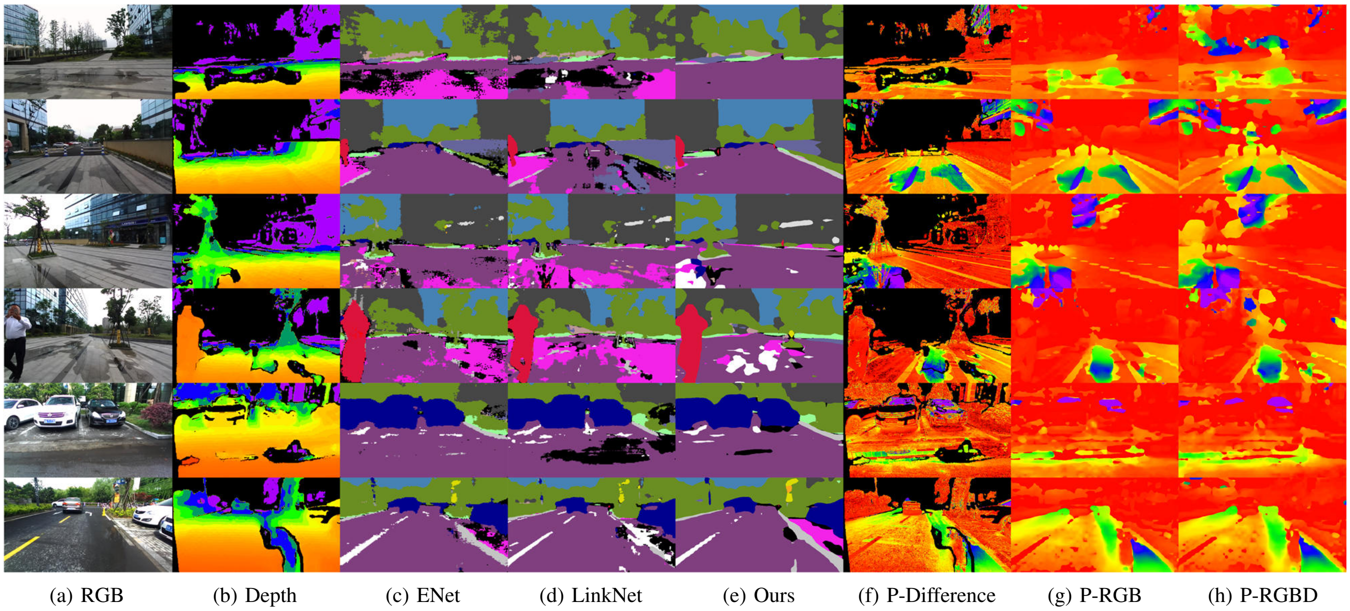


Fig. 5. Qualitative examples of polarization difference prediction beyond semantic segmentation on real-world images produced by our approach compared with ENet [19] and LinkNet [20]. From left to right: (a) RGB image, (b) Depth image, (c-e) Semantic segmentation produced by ENet, LinkNet and our ERF-PSPNet (sequential version), (f) Polarization difference captured with the polarized stereo camera [6], (g-h) Polarization difference predicted from RGB image and RGB-D image by our ERF-PSPNet (sequential version).

[5] F. Wang, S. Ainouz, C. Lian and A. Bensrhair, Multimodality semantic segmentation based on polarization and color images, *Neurocomputing*, 253, 2017, pp. 193-200.

[6] K. Yang, K. Wang, R. Cheng, W. Hu, X. Huang and J. Bai, Detecting traversable area and water hazards for the visually impaired with a pRGB-D sensor, *Sensors*, 17(8), 2017, p. 1890.

[7] K. Berger, R. Voorhies and L.H. Matthies, Depth from stereo polarization in specular scenes for urban robotics, In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 1966-1973.

[8] X. Huang, J. Bai, K. Wang, Q. Liu, Y. Luo, K. Yang and X. Zhang, Target enhanced 3D reconstruction based on polarization-coded structured light, *Optics Express*, 25(2), 2017, pp. 1173-1184.

[9] 4D Technology, Polarization camera for image enhancement, https://www.4dtechnology.com/products/polarimeters/polarcam/.

[10] D. Vorobiev, Z. Ninkov, N. Brock and R. West, On-sky performance evaluation and calibration of a polarization-sensitive focal plane array, In *Advances in Optical and Mechanical Technologies for Telescopes and Instrumentation II*, 2016, p. 99125X.

[11] J. Qi, C. He and D.S. Elson, Real time complete Stokes polarimetric imager based on a linear polarizer array camera for tissue polarimetric imaging, *Biomedical optics express*, 8(11), 2017, pp. 4933-4946.

[12] D. Wang and R. Liang, Simultaneous polarization Mirau interferometer based on pixelated polarization camera, *Optics letters*, 41(1), 2016, pp. 41-44.

[13] L. Yang, F. Tan, A. Li, Z. Cui, Y. Furukawa and P. Tan, Polarimetric Dense Monocular SLAM, In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3857-3866.

[14] M.P. Rowe, E.N. Pugh, J.S. Tyo and N. Engheta, Polarization-difference imaging: a biologically inspired technique for observation through scattering media, *Optics letters*, 20(6), 1995, pp. 608-610.

[15] J. Long, E. Shelhamer and T. Darrell, Fully convolutional networks for semantic segmentation, In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431-3440.

[16] F. Ma and S. Karaman, Sparse-to-dense: depth prediction from sparse depth samples and a single image, *arXiv preprint arXiv:1709.07492*, 2017.

[17] Y. Liu, X. Chen, H. Peng and Z. Wang, Multi-focus image fusion with a deep convolutional neural network, *Information Fusion*, 36, 2017, pp. 191-207.

[18] 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(12), 2017, pp. 2481-2495.

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

[20] A. Chaurasia and E. Culurciello, LinkNet: Exploiting Encoder Representations for Efficient Semantic Segmentation, In *2017 IEEE Visual Communications and Image Processing (VCIP)*, 2017, pp. 1-4.

[21] Q. Ning, J. Zhu and C. Chen, Very Fast Image Segmentation Using Hierarchical Dilation and Feature Refining, *Cognitive Computation*, 2018, 10(1), pp. 62-72.

[22] J. Kim, J. Baek, H. Choi and E. Kim, Wet area and puddle detection for Advanced Driver Assistance Systems (ADAS) using a stereo camera, *International Journal of Control, Automation and Systems*, 14(1), 2016, pp. 263-271.

[23] C.V. Nguyen, M. Milford and R. Mahony, 3D tracking of water hazards with polarized stereo cameras, In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 5251-5257.

[24] O. Miksik, V. Vineet, M. Lidegaard, R. Prasaath, M. Nießner, S. Golodetz, S.L. Hicks, P. Pérez, S. Izadi and P.H. Torr, The semantic paintbrush: Interactive 3d mapping and recognition in large outdoor spaces, In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 3317-3326.

[25] E. Schoop, J. Smith and B. Hartmann, HindSight: Enhancing Spatial Awareness by Sonifying Detected Objects in Real-Time 360-Degree Video, In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, p. 143.

[26] B. Mocanu, R. Tapu and T. Zaharia, When ultrasonic sensors and computer vision join forces for efficient obstacle detection and recognition, *Sensors*, 16(11), 2016, p. 1807.

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

[28] F. Yu and V. Koltun, Multi-scale context aggregation by dilated convolutions, *arXiv preprint arXiv:1511.07122*, 2015.

[29] H. Zhao, J. Shi, X. Qi, X. Wang and J. Jia, Pyramid scene parsing network, In *2017 IEEE Conference on Computer Vision and Patter Recognition (CVPR)*, 2017, pp. 6230-6339.

[30] G. Neuhold, T. Ollmann, S.R. Bulo and P. Kontschieder, The mapillary vistas dataset for semantic understanding of street scenes, In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5000-5009.

[31] X.Y.Z.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, *arXiv preprint arXiv:1711.01506*, 2017.