# PASS: Panoramic Annular Semantic Segmentation

Kailun Yang[1], Xinxin Hu[1], Luis M. Bergasa[2], Eduardo Romera[2], and Kaiwei Wang[1]

***Abstract*—Pixel-wise semantic segmentation is capable of unifying most of driving scene perception tasks, and has enabled striking progress in the context of navigation assistance, where an entire surrounding sensing is vital. However, current mainstream semantic segmenters are predominantly benchmarked against datasets featuring narrow Field of View (FoV), and a large part of vision-based intelligent vehicles use only a forward-facing camera. In this paper, we propose a Panoramic Annular Semantic Segmentation (PASS) framework to perceive the whole surrounding based on a compact panoramic annular lens system and an online panorama unfolding process. To facilitate the training of PASS models, we leverage conventional FoV imaging datasets, bypassing the efforts entailed to create fully dense panoramic annotations. To consistently exploit the rich contextual cues in the unfolded panorama, we adapt our real-time ERF-PSPNet to predict semantically meaningful feature maps in different segments, and fuse them to fulfill panoramic scene parsing. The innovation lies in the network adaptation to enable smooth and seamless segmentation, combined with an extended set of heterogeneous data augmentations to attain robustness in panoramic imagery. A comprehensive variety of experiments demonstrates the effectiveness for real-world surrounding perception in a single PASS, while the adaptation proposal is exceptionally positive for state-of-the-art efficient networks.**

***Index Terms*—Intelligent Vehicles, Scene Parsing, Semantic Segmentation, Scene Understanding, Panoramic Annular Images.**

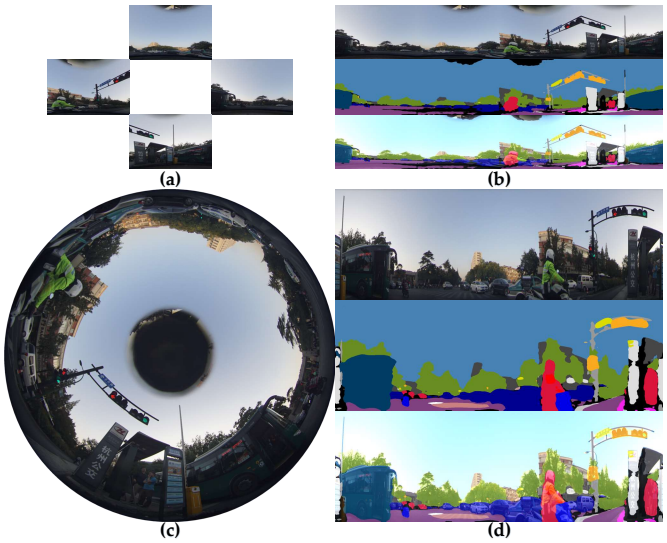



Fig. 1. (a)(b) Traditional semantic perception using surround pinhole views, (c) Raw panoramic annular image captured by our perception system with a single camera, (d) The proposed panoramic annular semantic segmentation on real-world surrounding view for 360° seamless scene understanding.

## I. Introduction

THE attracted attention of pixel-wise semantic segmentation in the context of Intelligent Transportation Systems (ITS) is rising, as most navigational perception tasks desired by autonomous vehicles and assisted ITS can be addressed in a unified manner [1][2], which traditionally rely on multiple detectors and expensive LiDAR/RADAR sensors that are typically employed in complex separate ways [3].

However, almost all semantic segmenters are benchmarked against conventional perspective images in existing datasets, such as Cityscapes [4] and Mapillary Vistas [5]. Besides, mainstream semantic perception frameworks are normally designed to work with vision sensors capturing a limited imaging angle such as standard forward-view pinhole cameras integrated on intelligent vehicles [6]. Nevertheless, pinhole imaging has a severe disadvantage: critical scene semantics move out of the Field of View (FoV). This renders semantic segmentation as an insufficient solution to automotive sensing and situational awareness, because autonomous navigation systems need measurably reliable and comprehensive perception of the entire surrounding, such that sufficient certainty can be propagated to upper-level applications. In this sense, extending semantic segmentation to panoramic perspective is vital for safe navigation, especially in metropolitan intersections with high traffic density and big volumes of information to be adequately handled. The associated question naturally emerges: can you parse the scene beyond the FoV [7]?

To this end, there were a few semantic perception platforms that have addressed panoramic segmentation by arraying several conventional cameras [6][8][9] or attaching fish-eye cameras with pronounced lens-introduced distortions [10][11][12][13][14][15]. However, these frameworks typically stitch segmented maps from multiple cameras with varying orientations [8][9], still only facilitating less than 180° semantic understanding of the forward surroundings [6][13], frequently resulting in inconsistent panoramic segmentation (see Fig. 1(a)(b)) or bird-eye interpretation that sacrifices safety-critical horizontal surrounding view above the horizon [11][14][15]. Notoriously, the number of devices that compose a perception system is one of the most critical parameters to be optimized, as deploying multiple cameras induces large latency and heavy computational burden, as well as the fulfillment of a set of hard tasks such as sensor calibration, synchronization and data fusion. Nevertheless, perception system using only a single camera to make prediction of 360°

Manuscript received May 1, 2019; revised June 21, 2019; accepted August 28, 2019. (Corresponding author: Kaiwei Wang.)

[1]K. Yang, X. Hu, and K. Wang are with State Key Laboratory of Modern Optical Instrumentation, Zhejiang Univeristy, 310027 Hangzhou, China (e-mail: elnino@zju.edu.cn; hxx_zju@zju.edu.cn; wangkaiwei@zju.edu.cn).

[2]L. M. Bergasa and E. Romera are with Department of Electronics, University of Alcalá, 28805 Madrid, Spain (e-mail: luism.bergasa@uah.es; eduardo.romera@uah.es).

panoramic per-pixel semantics in one pass is scarce in the state of the art.

In this paper, we bridge the gap by proposing a Panoramic Annular Semantic Segmentation (PASS) framework to "PASS beyond the FoV" using previously designed Panoramic Annular Lens (PAL) [16], whose compactness is a certainly desirable trait for diverse navigational applications [1][6]. Another tractable aspect is the distortion that has been well maintained in less than 1% [16], and the imaging model follows a clear f-theta law, convincingly appealing for deploying a panoramic semantic perception system characterized with comprehensiveness and wide applicability.

On the other hand, the fundamental challenge to accomplish this goal lies in the preparation of pixel-accurate annotations which is extremely labor-intensive and time-consuming. While state-of-the-art segmentation models demonstrate excellent performance by training with huge quantities of finely labeled images [17][18], applying this protocol to panoramic imagery is problematic, as it is hardly affordable to repeat the annotation procedure for all different conditions to have the same amount of high-quality data [4][5]. Instead, departing from the paradigm, if we could exploit conventional perspective images for training a panoramic semantic segmenter, it would be immensely beneficial for our omnidirectional sensing system to cover a comprehensive variety of driving/navigating situations, by taking advantage of the wealth of already openly available datasets.

More precisely, to yield PASS models for a holistic semantic understanding of the surrounding scene, we leverage large-scale databases like Vistas [5], bypassing the effort needed to create dense pixel-exact annotations. To preserve the contextual priors in the panoramic content after image unfolding, we propose a cluster of network adaptation techniques with our ERF-PSPNet [1][2] to infer semantically meaningful feature maps and fuse them to complete the panoramic segmentation (see Fig. 1(c)(d)) through the last fully convolutional layers. To improve the robustness, we apply a heterogeneous set of data augmentation methods, earning specialized knowledge and generalization gains in panoramic content with a view to real deployment. This paper is the extension of our conference paper [7], which has been extended with a detailed description of the proposed PASS framework, along with an extended set of experiments. Accompanying the framework, we provide the community with a PASS dataset to benchmark panoramic perception algorithms. The full dataset and corresponding PyTorch code of our framework are open-sourced at[1].

## II. Related Works

### A. Semantic Segmentation and Scene Parsing beyond the FoV

Semantic segmentation has progressed exponentially thanks to the breakthrough of Convolutional Neural Networks (CNNs). Fully Convolutional Network (FCN) [19], DeConvNet [20], UNet [21] and SegNet [22] represent the pioneering works. Their accuracies were surpassed by subsequent top-performing networks including PSPNet [18], FRRN [23], RefineNet [24], DeepLab [25], DenseASPP [26], PAN [27], OCNet [28] and ACNet [29] which can yield highly qualified and finely grained segmentation maps by using sophisticated network structures. Dilated convolution [30], pyramid pooling [18], atrous spatial pyramid pooling [25][26], context encoding [31] and object context pooling [28] introduced different senses of context and helped to push forward the performance boundary ceaselessly. Inevitably, with increased computational complexity, significant inference slow-down has been incurred, disqualifying these networks in real-time applications.

Another cluster of research efforts has been dedicated to addressing the restrictions of applications such as autonomous navigation, where real-time setups and light-weight segmenters are clearly preferred. To name a few of representative efficient networks, ENet [32], LinkNet [33], SQNet [34], ICNet [35], ESPNet [36], EDANet [37], BiSeNet [38], CGNet [39] and RPNet [40]. In previous works, we propose ERFNet [17] and ERF-PSPNet [1][2], which can perform semantic segmentation both efficiently and accurately, suitable for countless navigational applications. As a follow-up work, we extended dilated convolution [30] to hierarchical structures and predicted per-pixel polarization cues beyond semantic segmentation [41]. However, the comprehensiveness could be further improved as these research directions pertain to work with conventional pinhole cameras whose FoV is severely limited. Recently omnidirectional vision sensors have been increasingly attracting interests, as larger FoV of surrounding scenes can be captured. Nonetheless, contemporary works using omnidirectional cameras have predominantly focused on depth estimation [42] or visual localization [43][44].

In contrast, panoramic semantic segmentation, which has not been explicitly investigated, should be traced back to fish-eye image parsing. L. Deng *et al.* [10] overlapped pyramidal pooling [18] of encoded featured maps for fish-eye segmentation that theoretically facilitates the entire understanding of frontal hemispheric view. Á. Sáez *et al.* [12] followed this trend by implementing real-time semantic parsing for fish-eye urban driving images, and outperformed the seminal work [10] in terms of both inference speed and segmentation accuracy using ERFNet [17]. These two groups both extended their respective conference work. Á. Sáez *et al.* [13] validated against real fish-eye data with the eventual goal to complement a LiDAR sensor, where both sensors were equipped on their open-sourced electric car for assisting the elderly with driving tasks. In contrast, L. Deng *et al.* [11] used four wide-angle cameras to build a surrounding view whose semantic labels were projected to the accumulated 3D point cloud [45], although the original assumption was that only two cameras would be hypothetically needed to cover the 360° [10][12]. Keeping up with this trend, T. Sämann *et al.* [14] accelerated the forward pass of ENet [32] by using a channel pruning method, and enabled semantic bird-view interpretation with images from four raw fish-eye cameras. Y. Wu *et al.* [15] focused on parking lot and lane markings segmentation on panoramic surround view, which also relied on undistortion, warping and stitching operations by using four original fish-eye cameras.

[1]PASS: https://github.com/elnino9ykl/PASS

W. Zhou *et al.* [6] replaced a 56° FoV camera with three 100° FoV lens in an array, aiming to parse a full forward-facing panorama by stitching the undistorted fish-eye image segmentation maps. However, it was only able of perceiving the surroundings in front of the vehicle. R. Varga *et al.* [8] enforced panoramic automotive sensing with a super-sensor, whose images were segmented and unwarped on cylindrical projection surfaces. In spite of being able to attain horizontal 360° coverage of the vehicle surrounding, a large portion of vertical FoV was sacrificed to preserve straight lines. Similarly, K. Narioka *et al.* [9] pursued perception wideness by installing five cameras equiangularly on top of an instrumented car. They used a light-weight varient of SegNet [22] to be executable with sufficient speed. Experimentally, they found that there is an accuracy downgrade when using only forward-view camera images, which indicates that side-view knowledge are also critical and diverse viewpoints should be incorporated for training a panoramic semantic segmenter. Very recently, B. Pan *et al.* [46] aggregated the first-view observations from different angles to parse into a top-down-view semantic map for a deeper awareness of the surroundings with a better sensing of the spatial configurations. Our work differs fundamentally from these methods. Instead of being modeled in complex separate ways, we aim to use a single camera to comprehensively parse 360° real-world scenes in an efficient coherent manner, as PASS is made in a single pass.

### B. Semantic Segmentation Datasets and Data Augmentations

Semantic segmentation datasets have played an essential role and spurred key creativity in ITS research field. In the last years, numerous autonomous driving-oriented large-scale datasets have emerged such as Cityscapes [4], CamVid [47], BDD [48], IDD [49] and Mapillary Vistas [5]. Cityscapes is one of the milestone benchmarks with videos taken from a camera behind the windshield of intelligent vehicles. Being orders of magnitude larger than state-of-the-art datasets in terms of annotated frames, Apolloscape [50] specifically included some extremely cluttered and dynamic scenarios. WildDash [51] embraced the global diversity of traffic situations by incorporating test cases from all over the world. They extracted images from dashcam video materials, and created a checklist of hazards with the corresponding evaluation-oriented dataset. DarkZurich [52] as another evaluation dataset, was recorded from daytime, through twilight time to full nighttime. Nevertheless, these datasets only support forward-view of semantic scene understanding. Vistas is a crowd-sourced dataset that also attains global geographic reach of observations from different continents, and more appealingly it extends front-facing perspective to diverse viewpoints (*e.g.,* from roadways, sidewalks, unconstrained environments and off-road views), with promising implications in a broad variety of robotic vision applications [41].

These datasts have thousands of images, but even their diversity does not assure a good performance of current segmenters in unseen domains. To have more annotated data, Synthia [53] was proposed to facilitate learning with synthetic images. Its virtual acquisition platform consists of four 100° FoV binocular cameras with certain overlapping that can be used to create an omnidirectional view as validated in [11]. The TorontoCity benchmark [54] collected spherical panoramas from both drones and vehicles with pixel-level roadway annotations. WoodScape [55] comprises of four fish-eye cameras to observe the full surrounding of an automobile with semantically annotated images originating from distinct geographical locations. While these outdoor panoramic datasets are serviceable, severe distortions were introduced which are not compatible with images captured by our PAL system. In this work, like WildDash and DarkZurich, we also provide the research community with an evaluation dataset accompanying the proposed PASS framework, which features panoramic annular perspective and includes challenging frames such as cluttered scenes and hazy weather conditions to assess real-world reliability.

Under the vital topic of robust scene understanding, data augmentations have been broadly adopted to expand the datasets and combat over-fitting as an implicit regularization technique. To adapt to new, unseen domains, recent attempts have been made in [56][57], where augmentations were separated between geometry and texture, achieving desired network calibration and robustified traversability detection across wearable RGB-D cameras. We further explored how to make semantic segmentation work reliably in adverse conditions such as the nighttime [58][59]. J. Muñoz-Bulnes *et al.* [60] also indicated the path to attain enhanced learning generalization especially for road segmentation in bird-eye view, by randomly augmenting training data with geometric transformations and pixel-wise variations. S. Liu *et al.* [61] explored realistic image generation to balance the semantic label distribution by using Generative Adversarial Networks (GANs), and reported that the proportion of supplementary data in the training dataset should be well controlled. Regarding specialized augmentation methods to earn robustness against distortions that are inevitable when cameras have very large FoV as pointed out in [8], skew and gamma corrections were investigated in [6], while zooming policy was designed for fish-eye images in [10]. In particular, G. Blott *et al.* [62] proposed to depart from the central projection model [10] and used a large number of degrees of freedom for augmentation to model diverse camera orientations. In this work, we extend the zoom alteration and combine it with style transfer-based augmentation for panoramic semantic segmentation. Our systematic set of experiments, separating data augmentations between traditional, distortional and stylizational transformations, throws insightful hints on how robustness gains are earned across conventional/omnidirectional perspectives.

## III. PASS: Proposed Framework

### A. Training Stage

The overview of the proposed PASS framework is depicted in Fig. 2. In the training course, our publicly available semantic segmentation network ERF-PSPNet [1][2] is adapted, which is built using an efficient encoder from ERFNet [17] and a pyramid pooling-based decoder from PSPNet [18]. Our ERF-PSPNet inherits both technical gists including spatial factorized filters, sequential dilated convolutions and pyramidal

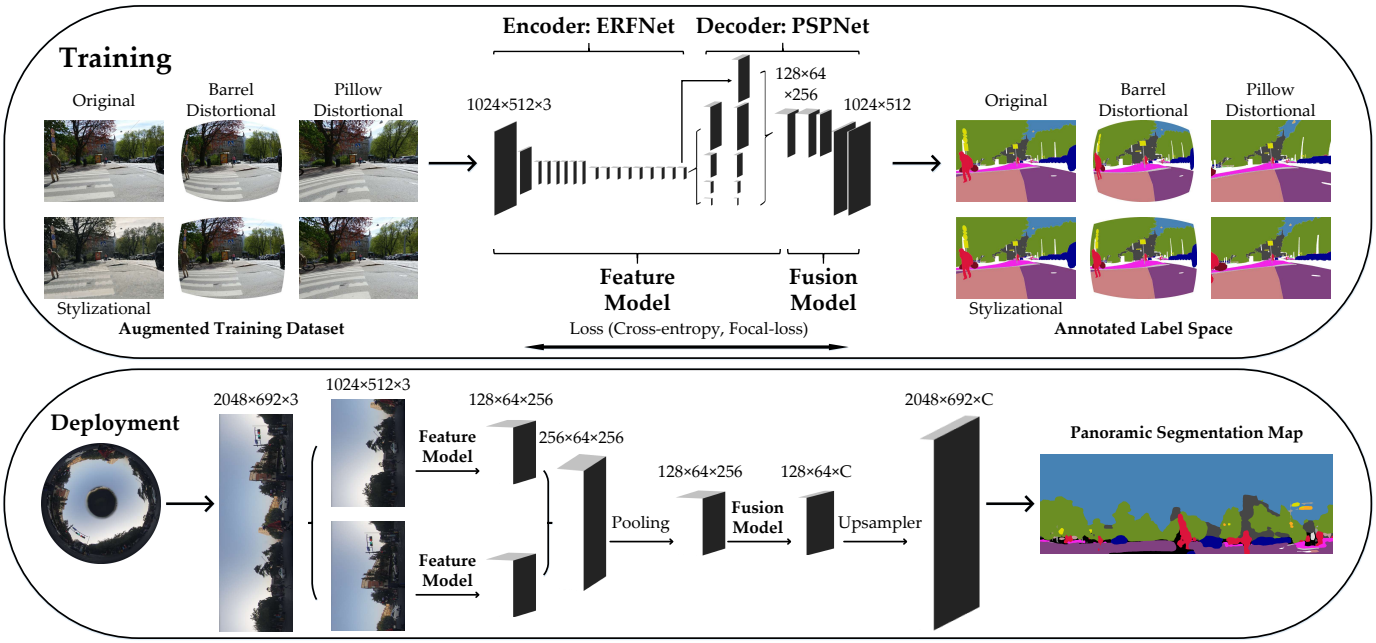



Fig. 2. The proposed panoramic annular semantic segmentation framework.

pooled representations, so as to strike an essential balance between real-time speed and segmentation accuracy, and outperforms ERFNet in context-critical domains [1]. Based on this architecture, semantic scene parsing can be addressed in a both efficient and accurate way, while the compactness is maintained to be easily deployed in an embedded system. By training on a conventional FoV imaging dataset, a light-weight segmentation model $F$ is yielded. Given a conventional FoV image, $I_c^{H\times W}$, a segmentation map, $S_c^{H\times W}$, at the inputting size $H\times W$, can be precisely predicted by $F$ that can also be separated into a feature model $F_e$ and a fusion model $F_u$, formally:

$$S_c^{H\times W} = F\left(I_c^{H\times W}\right) = F_u\left[F_e\left(I_c^{H\times W}\right)\right] \tag{1}$$

In this work, we re-purpose the ERF-PSPNet model $F$ and methodically adapt it in a way suitable for addressing panorama segments semantic segmentation, where global contextual information is rich and should be exploited in a deeper way than learning from local textures.

### B. Deployment Stage

In the deployment phase, the PAL system is calibrated and the panoramic annular image is unfolded using the interface provided by the omnidirectional camera toolbox [63]. The unfolded panorama is partitioned into $M$ segments as it is depicted in the following equation:

$$I_p^{H_p\times W_p} = \biguplus_{i=1}^{M}(I_i^{H_p\times \frac{W_p}{M}}) \tag{2}$$

where $I_p$ denotes the unfolded panoramic image whose size is $H_p\times W_p$, and $I_i$ denotes panorama segment.

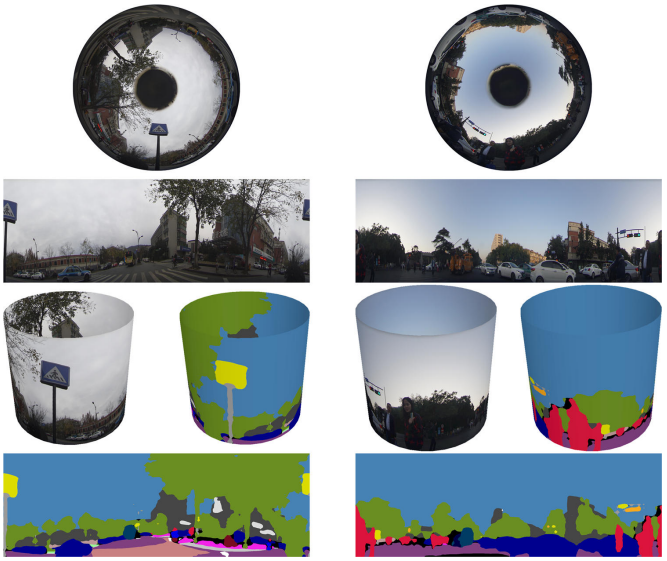
Vitally, in the re-separated ERF-PSPNet ($F_e + F_u$), the feature model $F_e$ is responsible for predicting high-level semantically meaningful feature maps of panorama segments and the fusion model $F_u$ is in charge of final classification and completing the full segmentation. To complete the panoramic parsing, a straightforward solution is to directly integrate the inferred pixel-wise probability maps of $M$ segments along the unfolding direction. Although it can instantly fulfill the segmentation, unsatisfactory discontinuity will be incurred for the loss of local context around the boundaries of neighboring segments. Instead, we propose to use only the feature model $F_e$, as shown in Fig. 2, which excludes the last convolution layer of ERF-PSPNet to predict feature maps of each segment ($I_i^{H_p\times \frac{W_p}{M}}$) taking into account there is a correspondence between features inferred from the panorama segments and features inferred from the conventional narrow FoV images used in the training:

$$F\left(\biguplus_{i=1}^{M}(I_i^{H_p\times \frac{W_p}{M}})\right) \equiv F\left(\biguplus_{j=1}^{N}(I_{c_j}^{H\times W})\right) \tag{3}$$

After the concatenation of the $M$ segments and a max-pooling process to recover the original feature model size, the entire panoramic annular image is smoothly parsed by the fusion model $F_u$, since semantically abstract features have already been extracted and aggregated. In other words, the output of the feature model already possesses all the necessary contextual information, hence the fusion model could work without much adaptation. Followed by a bilinear upsampler, the panorama segmentation map $S_p^{H_p\times W_p}$ is obtained by matching to the inputting size:

$$S_p^{H_p\times W_p} = F_u\left[\biguplus_{i=1}^{M} F_e\left(I_i^{H_p\times \frac{W_p}{M}}\right)\right] \tag{4}$$

where $\biguplus$ denotes concatenation of feature maps, which can also be considered as a feature denoising block to increase



(a) (b)

Fig. 3. Panoramic annular images can be folded back into 360° cylindrical rings for seamless padding and upsampling.

robustness and enforce smoothness when added at intermediate layers before the 1×1 convolution layer [64]. For illustrative purposes, $M$ is set to 2 in Fig. 2. In our experiment, different options ($M$=1, 2, ..., 6) are explored to study the effect of this FoV-related parameter on the final 360° segmentation performance.

### C. Network Adaptation

We further propose some network adaptation techniques to face borders discontinuity in the panorama (see Fig. 3) or overlapping of semantics across different segments when splitting the panorama, because impairing the context around the borders results in inconsistency and performance decrease. In the convolution layers, instead of traditional zero-padding around the feature map boundary, a column of padding is copied from the opposite border for both 3×3 and horizontal 3×1 convolution kernels, implementing continuity in the panorama. This is due to an unfolded panorama can be folded over itself by stitching the left and right borders together as depicted in Fig. 3. This operation was first introduced as ring-padding in [42] for monocular depth estimation without any quantitative validation. In this paper, we not only provide real-world accuracy analysis, but also extend this concept to factorized and dilated convolutions that are essential in state-of-the-art networks including our ERF-PSPNet for efficient aggregation of more contextual cues.

In our architecture, stacks of dilated convolution layers in the encoder of ERF-PSPNet help to exponentially enlarge the receptive field of convolution kernels [1][2]. Accordingly, the padding has been proportionally widened to the dilation rate. Moreover, we extend the ring-padding concept to the cross-segment padding case where the copy is made from the neighboring segment when partitioning the panorama into multiple segments. Furthermore, in the bilinear interpolation layers of our decoder, we include specialized ring-upsampling

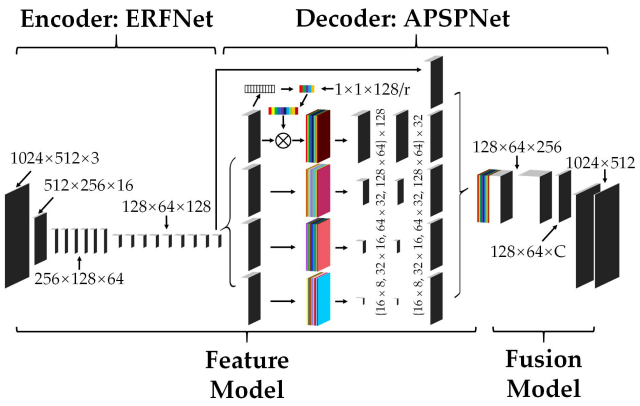



Fig. 4. The adapted ERF-APSP architecture.

and cross-segment upsampling to eliminate the undesirable boundary effects and enable true 360° forward passing.

Besides, as an alternative architecture, we adapt our ERF-PSPNet to ERF-APSPNet by introducing an Attention Pyramid Spatial Pooling (APSP) module in the decoder, while the encoder of ERFNet is kept in our architecture to maintain efficient inference, as illustrated in Fig. 4. Following the rationale that global contextual information is extremely important for panorama segments semantic segmentation as local textures are relatively distorted, we further spotlight the global context by re-calibrating channel-wise feature responses from the encoder, with the aim of increasing sensitivity to useful components and suppressing less informative features. Inspired by SENet [65], this is embodied by using global average pooling to squeeze spatial information into a channel descriptor, and element-wisely multiplying the feature map from the encoder and the corresponding re-shaped descriptor before each pyramid level of pooling. In this sense, the aggregated feature map after pyramid pooling contains more informative contextual information.

This channel attention operation is light-weight and enables adaptive feature refinement [65]. Based on this vital knowledge, we insert it in each pyramid level to exploit complementary context, where different responses are embedded and excited in this process as depicted in Fig. 4. In practice, we employ a multiplication layer to re-weight each level of features, and the reduction ratio $r$ is set to 16 to strike a trade-off between complexity and performance, boosting the accuracy with minimum sacrifices of inference speed, as APSP is located in the feature model which will be used for $M$ times for panoramic semantic segmentation.

### D. Data Augmentation

Our purpose is to learn from conventional FoV imaging dataset, while yielding models that must be robust against other domains and numerous blurs/distortions appear in unfolded panoramas. More precisely, Mapillary Vistas [5] is used for training, taking into a key consideration with respect to its high variance in camera viewpoint and focal length. Towards cross-domain robustness, different random data augmentation

techniques, separating in geometry, texture, distortion and style transformations, are performed in the training process:

*1) Traditional Geometric and Textual Data Augmentation:* Regarding geometric transformations, they are applied to both the original image and the ground truth mask. First, random rotations and shearings are implemented with degrees both uniformly sampled from the angles [-1°, 1°] to change positions of points while keeping lines straight. This augmentation is applied without cropping the original image, such that some black boundaries emerge, whose corresponding labels will not contribute to the loss function because the padding pixels are assigned to the "ignore label" of the classifier. Followed by rotation/shear augmentation, we implement translation and aspect-ratio augmentation. Although these transformations could be applied independently, cropping already has the benefits of translation [56], and better results could be obtained with combined scaling transformations [60]. In this sense, these augmentation effects are enabled together with scaling and cropping, by sampling distributions from [0.5, 1.0] to cut both the image height and width, and resize the randomly cropped sub-image to keep the same resolution in the feeding batch. Additionally, horizontal flipping (mirroring) is individually performed at a 50% opportunity to improve orientation invariance.

Regarding textual changes, brightness, contrast, saturation and hue variations are simultaneous augmented by selecting the values in random within the ranges [-0.1, 0.1] to improve the robustness against different illumination conditions and diverse color deviations.

*2) Extended Barrel and Pillow Distortion Augmentation:* To create synthetic distorted training samples from the Vistas dataset and extend the focal length data augmentation, it is important to refer to the projection model and the original alteration [10], where focal length $f$ was empirically set to map from each point $P_a = (x_a, y_a)$ in the augmented image to the conventional pinhole imaging point $P_c = (x_c, y_c)$ by adjusting the distance to the principal point $P_p$:

$$r_c = f \times tan(r_a/f) \tag{5}$$

where $r_c = \sqrt{(x_c - u_{cx})^2 + (y_c - u_{cy})^2}$ denotes the distance between the image point $P_c$ and the principal point $U_c = (u_{cx}, u_{cy})$ on the conventional image, while $r_a = \sqrt{(x_a - u_{ax})^2 + (y_a - u_{ay})^2}$ correspondingly denotes the distance between the image point $P_a$ and the principal point $U_a = (u_{ax}, u_{ay})$ on the augmented image. This mapping helps to add robustness against barrel distortion that is common in fish-eye images [10]. In this work, we extend the augmentation to address both barrel and pillow distortions by additionally creating training samples with adjusted distance:

$$r_c = f \times arctan(r_a/f) \tag{6}$$

The mapping relationship is determined by focal length $f$. Each image and its corresponding ground truth mask are transformed using the same mapping function to enable augmentation, but different interpolation methods are used: bilinear interpolation for images and nearest-neighbor interpolation for masks. This set of distortional transformations doesn't strictly follow the PAL imaging law, which could not be well modeled as under common cases the focal length parameters of conventional FoV imaging datasets are not available, but the joint use with geometric and textual augmentations helps to attain robustness to the distortions in panoramic content. This work adopts two scales of focal length ($f = 692$ or 1024) for both barrel and pillow distortion augmentations, whose augmentation effects can be seen in Fig. 2. Prior to this augmentation, the images from Vistas are homogenized to 2048×1384.

*3) Style Transfer Augmentation:* It is well known that large FoV imaging is generally associated with lower optical resolution [16]. The image resolutions of raw annular images (6000×4000) and unfolded panoramas (2048×692) are high, but the PASS imagery is also somewhat blurry compared with the high-definition VISTAS imagery, and a critical part of panoramic images are captured in hazy weather and low illumination conditions. To respond to these observations, resort is made to style transfer algorithms that have improved photo-realism very recently. More importantly, contemporary style transfer and image translation algorithms have lifted the requirement of paired samples from two different domains and could be utilized at the dataset level. To improve the robustness of semantic segmenters when taken out from their comfort zones to real-world non-ideal conditions, we leverage CycleGAN [66] to learn a transformation back and forth between the VISTAS and our PASS which are two unpaired imagery domains.

Instead of using solely stylized samples for training which is prone to over-fitting [61], we incorporate transformed training images from Vistas while preserving the original geometry of semantic labels as additional data, to jointly learn with original training images, while maintaining a suitable proportion of synthetic data according to [61]. In this way, the GAN-based transfer is re-purposed as a stylizational data augmentation technique to robustify against the blurs and compression artifacts present within panoramic imagery. Otherwise, the lack of invariance to blurring may bias the segmenter and corrupt the prediction when learning from total high-definition images.

## IV. Experiments

### A. Experiment Setup

The segmentation performance is evaluated on the Mapillary VISTAS validation dataset and our Panoramic testing dataset (PASS dataset), which is collected by using the previously designed compact PAL system that captures a FoV of 360°×75° (30°-105°), as shown in Fig. 5a. The PASS dataset contains 1050 raw and unfolded panoramic annular image pairs, from which 400 panoramas are finely labeled with masks on 4 critical street scene classes: *Car*, *Road*, *Crosswalk* and *Curb*, which are of immense significance for intersection perception and navigation assistance [67][68]. Compared to state-of-the-art hazard-aware evaluation datasets such as WildDash (70 public test cases) [51] and DarkZurich (20 labeled images) [52], our PASS (400 annotated panoramas) is a larger real-world dataset for assessing segmentation performance and measuring robustness. Schematically, four unfolded panoramas with ground-truth annotations are shown in Fig. 5b.

TABLE I
SEGMENTATION ACCURACY OF ERF-PSPNET ON MAPILLARY VISTAS DATASET.
***Pol***, ***StL***, ***Bil*** ETC. ARE ABBREVIATIONS OF THE CLASSES.

| **Pol** | **StL** | **Bil** | **TrL** | **Car** | **Tru** | **Bic** | **Mot** | **Bus** | **SiF** | **SiB** | **Roa** | **Sid** | **Cut** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 42.5% | 26.9% | 36.7% | 54.0% | 88.4% | 60.6% | 40.3% | 40.7% | 64.2% | 63.5% | 24.6% | 87.6% | 68.5% | 8.6% |
| **Pla** | **BiL** | **Cur** | **Fen** | **Wal** | **Bui** | **Per** | **Rid** | **Sky** | **Veg** | **Ter** | **Mar** | **Cro** | **mIoU** |
| 24.2% | 32.5% | 53.8% | 52.2% | 45.8% | 84.9% | 64.6% | 37.1% | 98.0% | 88.6% | 65.2% | 49.8% | 61.8% | 54.3% |

(a) (b)

Fig. 5. (a) Panoramic annular lens system; (b) Annotated example panoramas.

With the motivation of reflecting the robustness and real-world applicability, our dataset includes challenging scenarios, with a vital part of images captured at complex campuses/intersections in/around Zhejiang University at Hangzhou, China. Regarding the evaluation metrics, all numerical results are gathered by using the prevailing "Intersection over Union (IoU)" or "Pixel Accuracy (Acc)":

$$IoU = \frac{TP}{TP + FP + FN} \tag{7}$$

$$Acc = \frac{CCP}{LP} \tag{8}$$

where $TP$, $FP$ and $FN$ are respectively the number of true positives, false positives and false negatives at pixel level, while $CCP$ and $LP$ are respectively the number of correctly classified pixels and labeled pixels [69].

### *B. Training Details*

The Mapillary Vistas dataset [5] is utilized for training our semantic segmenter models to take advantage of its wide coverage and high variability in observation viewpoints, other than learning with only forward-view images [4][9]. Vistas is divided into 18000/2000/5000 images for training, validation and testing. Accordingly, we have 18000 training images from Vistas and its 2000 images for validation. Ground-truth labels of the 5000 testing images are not openly available. In this work, the annotated 400 panoramas from our PASS dataset are readily accessible for evaluation by pursuing the proposed deployment pipeline with the trained models.

Regarding the semantic categories, we use 27 out of the complete 66 classes to fit our campus/intersection scenarios and maintain the model efficiency. These 27 critical classes cover more than 95% of the labeled pixels, endowing the trained models with advanced capabilities to densely interpret metropolitan scenes. Regarding the CNN training setup, we train all our models under the same conditions using Adam optimization [70] with an original Learning Rate (LR) of $5\times10^{-4}$ and Weight Decay (WD) of $2\times10^{-4}$, exponentially decreasing LR until the loss converges with a maximum epoch number of 300 when feeding images at batch size of 6 and resolution of 1024×512. Two GPU cards including a NVIDIA GTX 1080Ti and a Titan RTX have been involved in the training of all the implemented networks. The 2D version of focal loss [71] with a focusing parameter $\gamma$=2 is adopted as the training criterion instead of conventional cross entropy. Following the scheme customized in [32], the class weights of loss function are determined as $w_{class} = \frac{1}{ln(c+p_{class})}$, where $c$ is set to 1.0005 to enforce the model to learn more about less frequent classes that have lower pixel probabilities $p_{class}$.

The CycleGAN is trained with the same hyperparameters as specified in [66] using 1050 images of the whole PASS dataset and 2000 images from the VISTAS validation dataset. To improve the generalization capacity, the encoder of our ERF-PSPNet and ERF-APSPNet is pre-trained on ImageNet [72] to seize the huge regularization opportunity afforded by knowledge transfer from larger and more diverse recognition datasets. When comparing against state-of-the-art semantic segmenters especially those efficiency-oriented models, we use the proposed parameters for them in their respective publications to ensure fair comparison, and re-separate them at comparable positions before 1×1 convolution layers as our networks.

Under these setups, our ERF-PSPNet reaches mean IoU (mIoU) of 54.3% on Mapillary Vistas validation dataset. This result achieved without any data augmentation is marked as the baseline, where per-class accuracy values are displayed in Table I, which verifies the learning capacity of our ERF-PSPNet on large-scale dataset.

### *C. On the Influence of Number of Segments*

Following the proposed segmentation pipeline, it is essential to study the effect of the number of segments ($M$) on the final performance in panoramic view. We experiment using diverse options of $M$ by partitioning the unfolded whole panorama into 1, 2, ..., 6 segments, corresponding to a FoV of 360°, 180°, ..., 60° per segment. As displayed in Table II, if only a single feature model ($M$=1) is used for the whole panorama, the context is too wide and results are clearly worse than

TABLE II
SEGMENTATION ACCURACY OF ERF-PSPNET ON PASS DATASET USING DIFFERENT NUMBER OF SEGMENTS. **BLUE** DENOTES HIGHER IOU WITH SPECIALIZED PADDING AND UPSAMPLING. **RED** HIGHLIGHTS THE BEST IOU.

| Number of Segments | FoV per Segment | Car | Road | Crosswalk | Curb |
|---|---|---|---|---|---|
| 1 | 360° | 71.8% **72.2%** | 65.7% **66.4%** | 29.2% **30.6%** | 18.4% 18.2% |
| 2 | 180° | 87.7% **88.2%** | **77.6% 78.8%** | 49.5% **50.4%** | 29.4% **30.3%** |
| 3 | 120° | 90.6% **91.0%** | 77.5% **78.3%** | **53.5% 53.9%** | 32.1% **32.8%** |
| 4 | 90° | **91.0% 91.4%** | 76.7% **77.6%** | 52.6% **52.9%** | **32.9% 33.4%** |
| 5 | 72° | 90.4% **90.7%** | 76.3% **76.8%** | 51.2% **51.6%** | 32.6% **33.0%** |
| 6 | 60° | 89.3% **89.6%** | 75.5% **75.9%** | 48.7% **49.2%** | 31.7% **32.5%** |

when the segment is more adapted to exploit the features of the classes. Consequently, the 360°-per-segment model suffers from an intolerably low accuracy with incompatible contextual cues when taken from conventional imagery to the omnidirectional imagery. In comparison, the 180°-per-segment predictor achieves the highest IoU on roadway segmentation, while the 120°-per-segment predictor outperforms others in terms of crosswalk segmentation. Smaller classes will require more segments than for the segmentation of cars and curbs, 90°-per-segment is the optimal option ($M$=4). Notably, regardless of segments number, the use of specialized padding and upsampling helps to boost the accuracy in almost all classes and options, demonstrating the effectiveness of our network adaptation proposal as one of the key enablers to fulfill accurate 360° semantic segmentation.

The segments finding is also consistent with the qualitative results. As comparably visualized in Fig. 6, the 360°-per-segment results are undesirable with limited detectable range of traversable areas, *e.g.*, roadways and sidewalks. In Fig. 6a, the 360°-per-segment model misses most pedestrians, and 180°-per-segment model only detects part of the person standing by the right sidewalk. Very intriguingly in Fig. 6b, the 360°-per-segment approach has classified both zebra crosswalks as general road markings, while the 180°-per-segment solution only correctly identifies a crosswalk region. One plausible hypothesis is that in most of the training samples, only one crosswalk region will be observed, hence 120°/90°-per-segment models are clearly better for crosswalk and sidewalk detection as well as the segmentation of diverse vehicles, pedestrians, riders and curbs. This also verifies that our ERF-PSPNet has successfully exploited such global contextual information, which is important for panorama segments semantic segmentation. On the other side, when using more feature models ($M$=5 or 6), the whole semantic panorama map tends to become fragmented. To maintain a good trade-off, we set $M$ to 4 in the following experiments.

This claim is also supported by testing with different FoVs of inputs in the panorama. As shown in Fig. 7, we crop different FoVs around the panorama center with variations of 10° for each point, and input the cropped images from the PASS dataset to the network. When inputting 90° images, our augmented ERF-PSPNet reaches the highest mIoU without any adaptation, verifying that $M$=4 is a suitable setting. This FoV implicitly corresponds to the focal length distribution of the Vistas dataset [5], which is mostly concentrated in the 25-35mm range (corresponding to a horizontal FoV of 54.4°-71.5°) along with a critical part of images taken by wide-angle cameras with focal length ranging between 15-20mm (84.0°-100.4°), as it involves diverse sensors for image capturing. Additionally, we find that when the inputting FoV is very narrow (*i.e.*, 10°-30°), the accuracy is even worse than the situation of using the whole 360° FoV panorama. Furthermore, when the inputting FoV exceeds 180°, the decline of the accuracy of crosswalk segmentation is the sharpest, illustrating that the crosswalk classification is highly related with global contextual information, which preliminarily supports the hypothesis regarding crosswalk segmentation and the training samples.

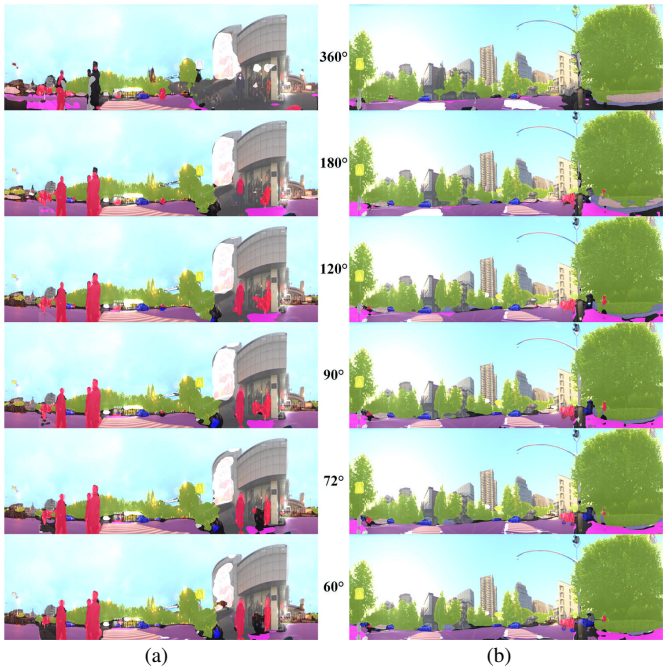



Fig. 6. Qualitative examples of semantically masked panoramic images by using our augmented PASS framework with different inference settings. From top to bottom: 360°-per-segment, 180°-per-segment, 120°-per-segment, 90°-per-segment, 72°-per-segment and 60°-per-segment results.

### D. On the Robustness of Panoramic Segmentation

In this experiment, taking an essential stride to delve into "accuracy" and "robustness", we analyze the gap between these two concepts under the topic of panoramic semantic segmentation. We collect the segmentation accuracy on the fully labeled Vistas validation dataset, in contrast with the real-world accuracy on PASS for testing (both in IoU), as displayed

TABLE III
ON THE ROBUSTNESS OF SEMANTIC SEGMENTATION ACROSS DOMAINS.

| Model | On VISTAS (Validation Dataset) | | | | | On PASS (Testing Dataset) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | mIoU | Car | Road | Crosswalk | Curb | Car | Road | Crosswalk | Curb |
| Baseline | **54.3%** | **88.4%** | **87.6%** | **61.8%** | **53.8%** | 86.1% | 71.6% | 40.2% | 32.8% |
| Distortional Augs (D) | 53.4% | 88.1% | 87.0% | 61.2% | 52.4% | 89.8% | 73.3% | 30.7% | 30.2% |
| Traditional Augs (T) | 52.9% | 87.6% | 86.6% | 61.7% | 49.0% | 90.4% | 74.2% | 41.1% | 33.2% |
| Combination (T+D) | 51.7% | 87.4% | 86.4% | 61.2% | 47.5% | 89.8% | 75.8% | 40.0% | 31.2% |
| Stylizational Augs (S) | 52.9% | 87.8% | 87.2% | 61.7% | 52.5% | 89.8% | 72.2% | 48.3% | 32.3% |
| All Augs (T+D+S) | 52.1% | 87.1% | 86.9% | 60.2% | 49.2% | **91.4%** | **77.6%** | **52.9%** | **33.4%** |

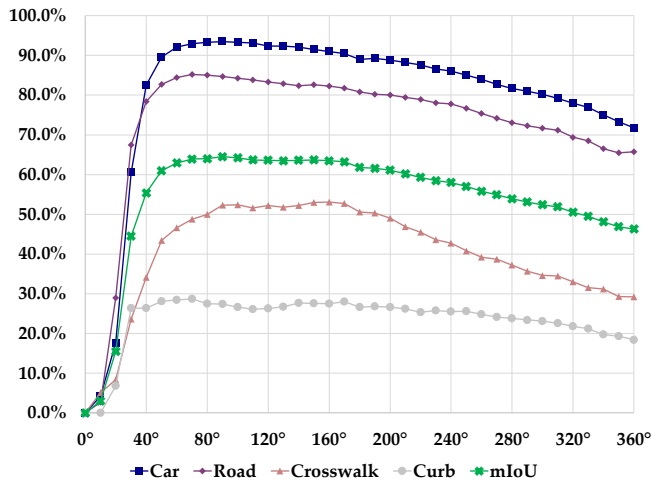
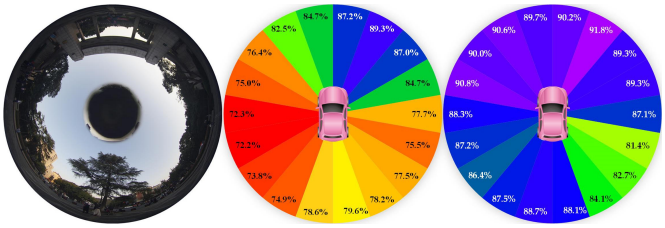



Fig. 7. The accuracy curves of ERF-PSPNet on different classes measured in IoU by using different FoVs of inputs in the panoramas.

in Table III. The proposed set of distortional (barrel+pillow) augmentations has incurred an accuracy downgrade on the validation dataset that does not contain distorted images. This is reasonable but we observe that on PASS dataset, the segmentation accuracy has been dramatically boosted in terms of cars and roadways, as normally roadways appear consistently in the lower part of the images across VISTAS/PASS domains and cars tend to be distorted in a uniform way, although it does not necessarily mean that all unseen data with crosswalks and curbs will face the modeled distortions. Noticeably, applying the traditional (geometry+texture) alterations also produces a large improvement, which makes sense since a certain part of intersections in the PASS dataset are not as highly illuminated as most scenarios in Vistas, needless to emphasize that the augmented aspect ratio is critical for panoramic segmentation. Traditional augmentation, being aggressive to attain high generalization capacity, also induces a slight accuracy decrease on validation dataset as the accuracy in the unseen panoramic imagery domain greatly increases, which further gives an intuition on how augmenting data highly prevents overfitting in the familiar domain, and helps yielding robust models for deployment in unseen, yet distinct domain. Based on this notion, we combine the distortional and traditional augmentations for training, and elevate to even higher accuracy of roadway segmentation without having seen any image from the panoramic imagery domain, which is one of the most important perception tasks within the context of autonomous navigation [1].

Regarding the effect of stylizational data augmentation by incorporating supplementary transferred images, it is noteworthy that the IoU of crosswalk segmentation has been improved to a great extent, which is due to that within panoramic imagery, most of the crosswalks are not as clear as those in Vistas dataset. The style converter excels exactly at generating realistic blurs with an example visualized in Fig. 2, thus making the augmented model more prepared against the panoramic imagery domain. Another key observation is that stylizational data augmentation only slightly improves the performance of road detection, which can be complemented by using distortional and traditional augmentations that have already reached high accuracy of roadway segmentation. When combining all heterogeneous data augmentations, the best accuracy boosts have been achieved for all classes, outperforming any independent augmentation by significant margins. This outstanding accuracy also demonstrates that robust panoramic segmentation is reachable against the challenging real-world PASS dataset. Based on such compelling evidence, one valuable insight gained from this cross-perspective experiment is that conceptually, the divergence of "accuracy" and "robustness" is not only a matter of CNN learning capacity, but also a matter of training sample diversity.

Fig. 8 also demonstrates the effectiveness of the proposed full set of data augmentation, which is a visualization of roadway segmentation Pixel Accuracy (Acc) values on the panoramic dataset by using the PASS model without/with data augmentation. Following [43], we partition all panoramic testing images into 18 directions, turning out that the augmented model improves a lot upon the baseline in all directions, while the advantage is also pronounced in forward-view directions, reaching accuracies of over/near 90.0% widely profitable for autonomous and assisted ITS systems. While the same panoramic view can be achieved from 4-6 cameras surrounding a vehicle, our system only uses a single camera with promise of good performances in certain safety-critical directions. Furthermore, Fig. 9 shows the montage of diverse semantic panorama maps of challenging frames in the PASS dataset. It can be easily noticed that in all qualitative segmentation examples of both campuses and complex intersections, our augmented model delivers impressive 360° semantic segmentation regardless of the distortions and blurs, owing to the proposed PASS framework and the extremely positive effects of the our network adaptation and data augmentation strategies in refining and robustifying panoramic segmentation.




(a) (b) (c)

Fig. 8. (a) A raw panoramic annular example image to indicate the orientation, (b) Segmentation accuracy values in different directions without data augmentation, and (c) with all data augmentations.

### *E. Comparison with the State of the Art*

To compare with state-of-the-art semantic segmentation networks, we contrast our ERF-PSPNet and ERF-APSPNet against known real-time models including ENet [32], LinkNet [33], SQNet [34], ICNet [35], ESPNet [36], EDANet [37], BiSeNet [38], CGNet [39] and our previous ERFNet [17]. As ERF-PSPNet is a rational combination of ERFNet and PSPNet, we also compare with PSPNet18 where efficient ResNet18 [73] is used as its backbone. There are top-performing networks like RefineNet [24] and DeepLab [25], but they are extremely inefficient, thus being computationally intensive to train on cost-effective GPUs and too heavy to deploy on embedded processors. Taking into account the computation constraints in intelligent vehicles and wearable robotics, we emphasize that our networks differ from those sophisticated models and mainly compare with light-weight networks for real-time semantic segmentation, shortening the re-training time of existing models. As displayed in Table IV and Fig. 10, we comprehensively test all networks on both VISTAS and PASS datasets in terms of mIoU and Frames Per Second (FPS). The FPS metric directly corresponds to the processing time on a single modern GPU NVIDIA Titan RTX, where the batch size has been set to 1 to simulate real-time applications. We report the mean FPS values over 400 forward passes running through all panoramas in the PASS testing dataset. Note that the mIoU scores are not comparable across datasets, as the number of classes and percent of annotated pixels are different, but PASS represents an unseen challenging real-world domain.

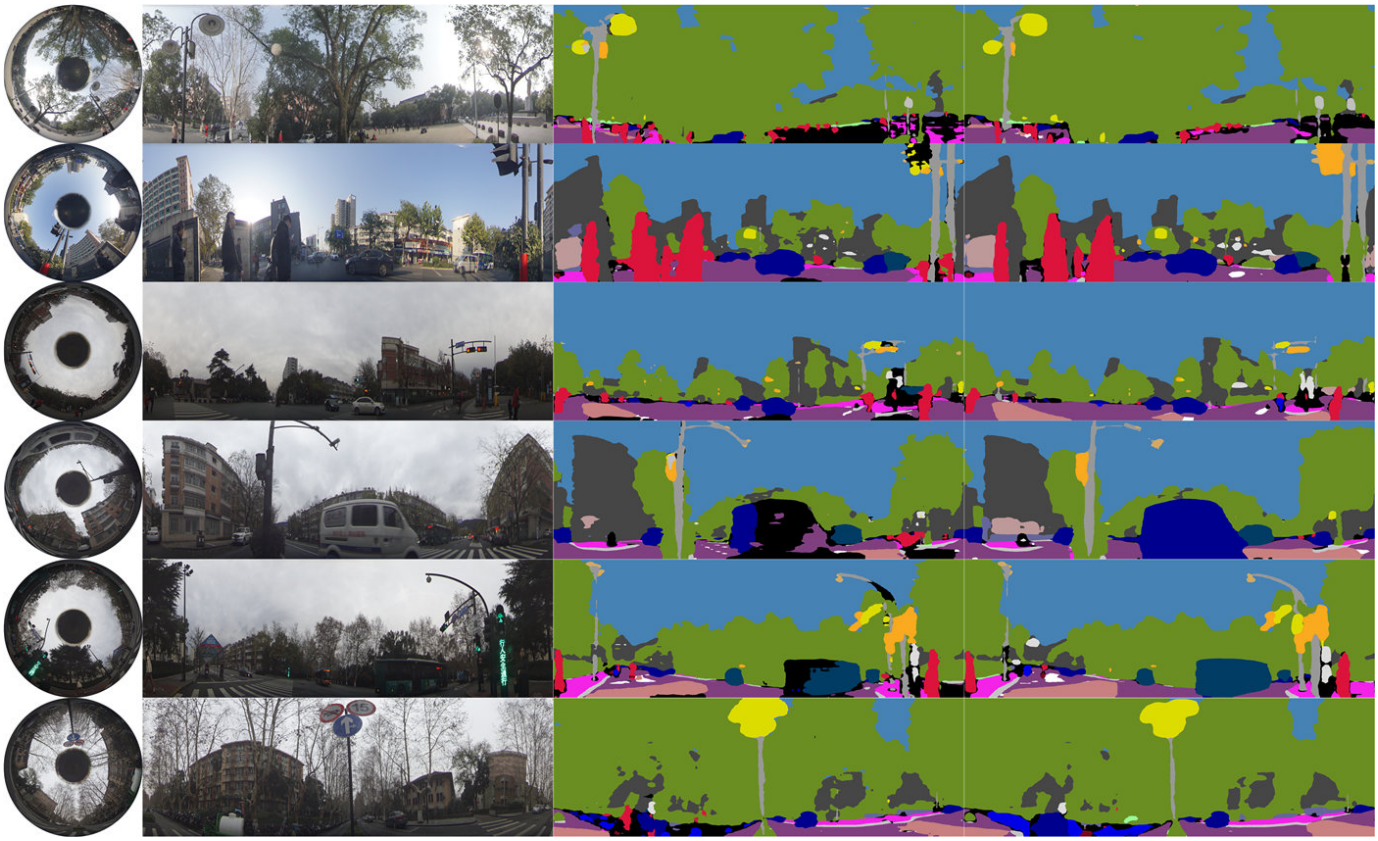
Table IV shows the accuracy and speed of all the networks on two datasets: VISTAS and PASS. An important question to ask is whether the proposed overall adaptation methodology enables reliably better performance than the baseline, *i.e.*, using the whole panorama to predict in an end-to-end way. To answer this question, we experiment with two settings: using only 1 feature model without any network adaptation; and using $M$ feature models with 360° network adaptation operations ($M$ has been set to 4). Overall, the network adaptation proposal is consistently effective for all the trained networks, as it can be easily seen that the mIoU values have been improved by significant margins, even more than 20.0%. In other words, the adaptation methodology is not strictly tied to a concrete architecture and can be easily deployed with other networks, which helps to yield promising models in panoramic imagery domain.

TABLE IV
COMPARISON WITH STATE-OF-THE-ART NETWORKS.
BLUE DENOTES THE ACCURACY TESTED WITHOUT MODEL ADAPTATION.
RED DENOTES THE ACCURACY TESTED WITH MODEL ADAPTATION.

| Network | On VISTAS | | On PASS | |
|---|---|---|---|---|
| | mIoU | FPS | mIoU | FPS |
| ENet [32] | 47.0% | 32.0 | 62.0% (37.4%) | 23.2 |
| LinkNet [33] | 47.7% | 185.7 | 58.6% (37.7%) | 75.1 |
| SQNet [34] | 39.5% | 178.0 | 50.1% (36.2%) | 71.3 |
| ICNet [35] | 43.1% | 191.3 | 46.9% (31.2%) | 79.0 |
| ESPNet [36] | 41.5% | 172.6 | 49.9% (30.1%) | 49.0 |
| EDANet [37] | 49.6% | 92.6 | 59.1% (36.5%) | 38.4 |
| BiSeNet [38] | 49.5% | 139.8 | 44.9% (35.7%) | 62.5 |
| CGNet [39] | 52.8% | 54.5 | 49.9% (34.1%) | 21.2 |
| ERFNet [17] | 52.7% | 77.8 | 62.9% (41.2%) | 34.7 |
| PSPNet18 [18] | 50.4% | 198.0 | 60.0% (40.8%) | 88.9 |
| **ERF-PSPNet** | **52.1%** | **108.7** | **63.8%** (**46.3%**) | **40.2** |
| **ERF-APSPNet** | **54.8%** | **96.0** | **64.4%** (**41.1%**) | **38.0** |

On PASS, our ERF-PSPNet and ERF-APSPNet reach the highest performance, while on VISTAS our ERF-APSPNet is the best. The channel-wise feature selection helps to gain a significant improvement of 2.7% over basic ERF-PSPNet on VISTAS, which is more notable than the improvement on PASS which mainly contains frequent classes for testing, as the attention operation is more advantageous for context-critical less frequent classes, while computational efforts only increase slightly. However, we observe that when testing without network adaptation, the IoU of ERF-APSPNet is lower than ERF-PSPNet, which is because that the attention module is exceptionally beneficial for exploiting global contextual information, thus impairing the performance when using the whole panorama for prediction where global context is incompatible. This is consistent in the cases of CGNet and BiSeNet as they also use attention modules in their architectures, especially CGNet that sequentially stacks a large number of its context guided blocks to achieve a very good accuracy on VISTAS but works unsatisfactorily if taken to the panoramic imagery without using our network adaptation strategy. In these senses, the suitability of our pyramid attention module is confirmed. Importantly, this result further verifies our hypothesis regarding the influence of number of segments and the context compatibility issue, backing up that the overall network adaptation methodology is critical to take advantage of wealthy knowledge learned from conventional FoV imaging dataset.

Regarding ESPNet and ENet, they suffer from limited learning capacity, thus being not accurate enough in both domains. Our ERF-PSPNet and ERF-APSPNet are both more efficient and more accurate than ERFNet and ENet on PASS, although ENet is one of the fastest segmentation network in the state of the art. We find that with adequate computational resources, our networks with fewer layers in max depth are faster than ENet, where similar results were reported in [40]. Qualitatively, as shown in Fig. 11, our ERF-PSPNet is more robust than ENet in the PASS imagery domain, especially evident for large objects and less frequent semantic classes as our networks have larger learning capacity and wider effective receptive field. ENet drops the last stages of the model,



(a) Raw (b) Unfolded panorama (c) Baseline (no augments) (d) All-augments PASS

Fig. 9. Qualitative examples of panoramic annular semantic segmentation: (a) Raw panoramic annular images, (b) Unfolded panoramic images, (c) Segmentation maps without data augmentation, and (d) with all data augmentations.

resulting in a too small receptive field to correctly classify large objects, such as the buses. In contrast, our networks maximize the usage of multi-scale context representations. Regarding the speed, networks based on ResNet18 are both very fast on Titan RTX such as PSPNet18, ICNet and BiSeNet, while our networks also realize above real-time predictions on both VISTAS and PASS datasets.

Fig. 10 visualizes the comparison. In Fig. 10a, it can be easily seen that our ERF-PSPNet and ERF-APSPNet have both achieved very high accuracies when facing the large-scale VISTAS and the unseen PASS domains. This outstanding result demonstrates the generalization capacity and high robustness of our networks, qualifying the usage for a wide spectrum of driving and navigating conditions. In Fig. 10b, it can be observed that our networks have achieved good trade-offs between accuracy and efficiency on PASS, reaching highest mIoU scores and above real-time inference speeds, which make them ideally suitable for the challenging real-world 360° semantic segmentation.

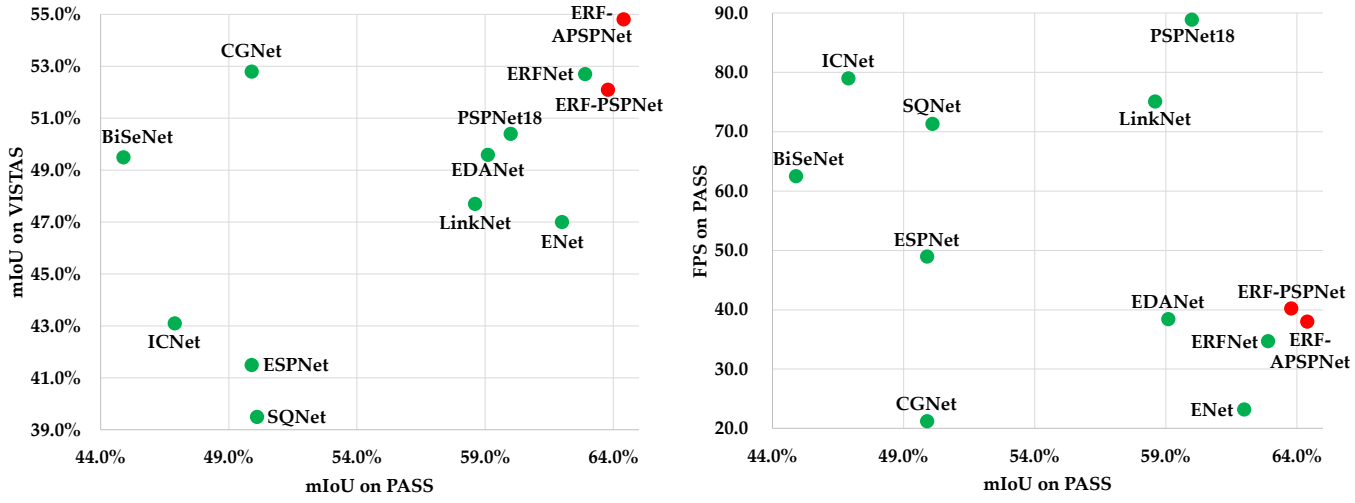
We additionally analyze the speed of PASS on embedded GPU processors including NVIDIA Jetson Nano and TX2, as shown in Table V. The proposed ERF-PSPNet and ERF-APSPNet are tested in two settings: viewing the whole panorama as 1 segment without any adaptation; and using 4 segments along with the best-performing adaptation strategy. There is a trade-off between accuracy and efficiency regarding the number of segments. Using 4 feature models is the most accurate setting if the computational budget is available, but using 1 feature model can still work. While the adaptation strategy trades efficiency for accuracy, there are ways to increase performance with no speed penalty. One of those ways is to view the 4 panorama segments as a simultaneously fed-in batch, but it neglects the warp-around connections which will create blind spots near the image/segment borders where we are unable to seamlessly interpret the environment. Furthermore, on the extremely portable embedded GPU Nano, PASS remains fast. At the resolution of 1024×512, a single ERF-PSPNet forward pass reaches more than 15FPS that is promising to be integrated in autonomous vehicles and assistive mobility systems.

TABLE V
SPEED ANALYSIS OF PASS (MEASURED IN FPS).

| GPU Processor | Nano | | TX2 | | Titan RTX | |
|---|---|---|---|---|---|---|
| Number of Segments | 1 | 4 | 1 | 4 | 1 | 4 |
| ERF-PSPNet | 15.9 | 4.2 | 17.2 | 6.6 | 108.7 | 40.2 |
| ERF-APSPNet | 12.8 | 3.5 | 14.3 | 5.1 | 96.0 | 38.0 |

## V. CONCLUSION AND FUTURE WORKS

In this paper, we look into the expansion of the Field of View of perception platforms by proposing a Panoramic Annular Semantic Segmentation (PASS) framework that promisingly endows automated intelligent vehicles or assisted naviga-




(a) (b)

Fig. 10. (a) Comparison of accuracies in mIoU on VISTAS and PASS datasets, (b) Comparison of accuracy in mIoU and speed in FPS on PASS dataset.

tion systems with advanced capabilities to accurately interpret the surroundings in a universal and comprehensive manner. Our approach enables fully dense and seamlessly panoramic semantic segmentation, meanwhile leaving opportunities open to fuse with LiDAR and RGB-D point clouds that could be displaced to lower priorities due to the prohibitive costs of those sensors.

With a new real-world evaluation dataset, the extensive set of experiments demonstrates that across domains, the robustness of 360° scene understanding has been augmented, even in complex metropolitan campus/intersection scenarios with a great deal of clutter and high traffic density. A comprehensive variety of comparison also verifies the effectiveness of the proposed network adaptation strategy for all the trained efficient state-of-the-art models, while our ERF-PSPNet and ERF-APSPNet achieve highest accuracies and above real-time parsing speeds in the panoramic imagery.

In the future, we have the intention to explore more network architectures, segmentation pipelines and augmentation strategies. Particularly, we are interested in non-local and detail-sensitive modules to further optimize the trade-off between prediction accuracy and efficiency. We aim to deploy PASS on real instrumented vehicles and mobile robotics, and investigate the benefit of panoramic semantics for upstream navigation components like visual odometry systems. In addition, the PASS dataset will be expanded by labeling more classes such as pedestrians and sidewalks.

## ACKNOWLEDGMENT

This work has 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 funded in part from the Spanish MINECO/FEDER through the SmartElderlyCar project (TRA2015-70501-C2-1-R) and from the RoboCity2030-DIH-CM project (P2018/NMT-4331), funded by Programas de actividades I+D (CAM) and cofunded by EU Structural Funds.

## REFERENCES

[1] K. Yang, L. M. Bergasa, E. Romera, R. Cheng, T. Chen, and K. Wang, "Unifying terrain awareness through real-time semantic segmentation," in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 1033–1038.

[2] K. Yang, K. Wang, L. M. Bergasa, E. Romera, W. Hu, D. Sun, J. Sun, R. Cheng, T. Chen, and E. López, "Unifying terrain awareness for the visually impaired through real-time semantic segmentation," *Sensors*, vol. 18, no. 5, p. 1506, 2018.

[3] N. Long, K. Wang, R. Cheng, K. Yang, and J. Bai, "Fusion of millimeter wave radar and rgb-depth sensors for assisted navigation of the visually impaired," in *Millimetre Wave and Terahertz Sensors and Technology XI*, vol. 10800. International Society for Optics and Photonics, 2018, p. 1080006.

[4] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pp. 3213–3223.

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

[6] W. Zhou, A. Zyner, S. Worrall, and E. Nebot, "Adapting semantic segmentation models for changes in illumination and camera perspective," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 461–468, 2019.

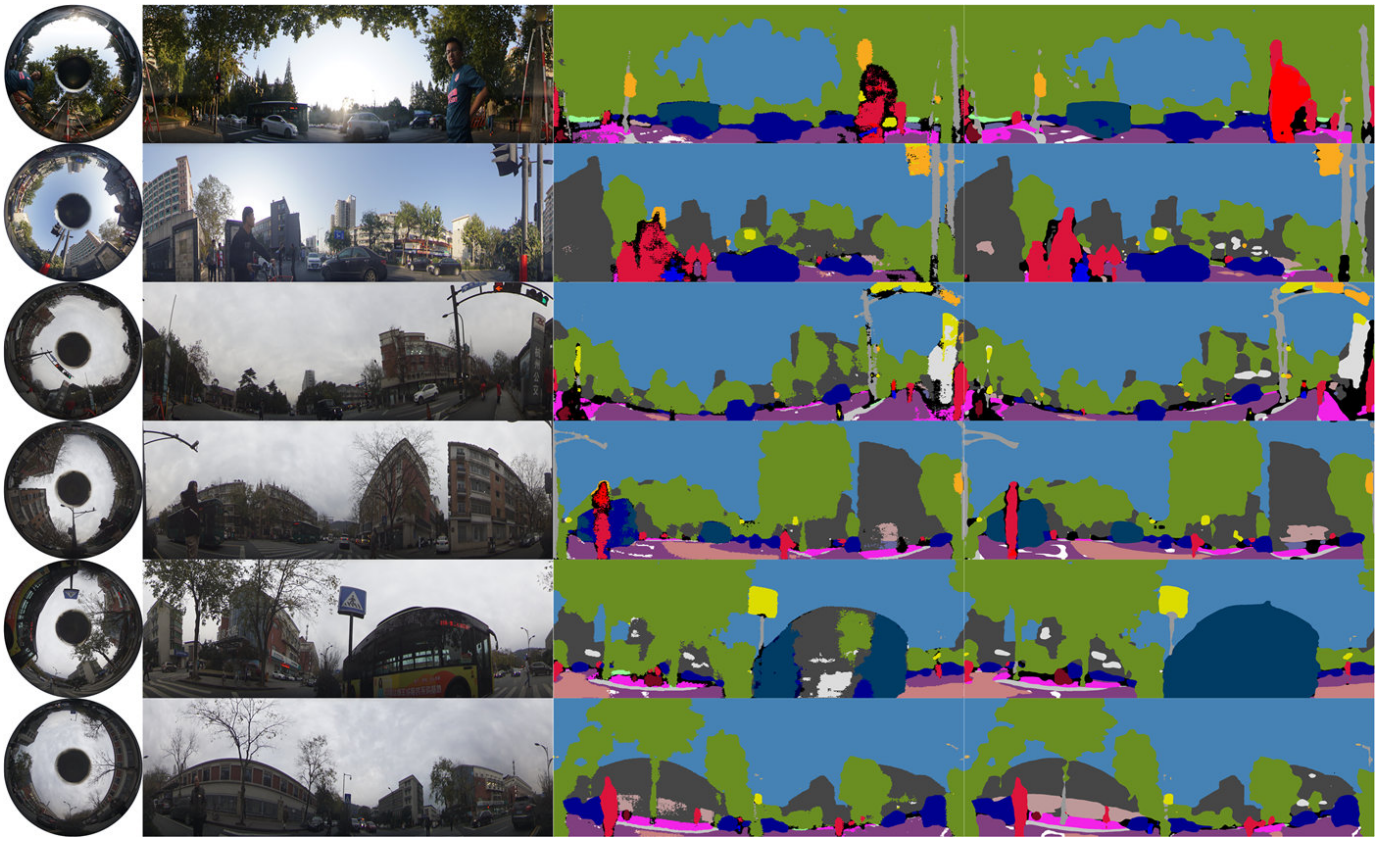
[7] K. Yang, X. Hu, L. M. Bergasa, E. Romera, X. Huang, D. Sun, and K. Wang, "Can we pass beyond the field of view? panoramic annular semantic segmentation for real-world surrounding perception," in *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2019, pp. 374–381.

[8] R. Varga, A. Costea, H. Florea, I. Giosan, and S. Nedevschi, "Super-sensor for 360-degree environment perception: Point cloud segmentation using image features," in *Intelligent Transportation Systems (ITSC), 2017 IEEE 20th International Conference on*. IEEE, 2017, pp. 1–8.

[9] K. Narioka, H. Nishimura, T. Itamochi, and T. Inomata, "Understanding 3d semantic structure around the vehicle with monocular cameras," in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 132–137.

[10] L. Deng, M. Yang, Y. Qian, C. Wang, and B. Wang, "Cnn based semantic segmentation for urban traffic scenes using fisheye camera," in *Intelligent Vehicles Symposium (IV), 2017 IEEE*. IEEE, 2017, pp. 231–236.

[11] L. Deng, M. Yang, H. Li, T. Li, B. Hu, and C. Wang, "Restricted deformable convolution based road scene semantic segmentation using surround view cameras," *arXiv preprint arXiv:1801.00708*, 2018.



(a) Raw (b) Unfolded panorama (c) ENet [32] (d) Ours

Fig. 11. Qualitative examples of panoramic annular semantic segmentation: (a) Raw panoramic annular images, (b) Unfolded panoramic images, (c) ENet [32] with data augmentation, and (d) Our ERF-PSPNet with data augmentation.


[12] Á. Sáez, L. M. Bergasa, E. Romera, E. López, R. Barea, and R. Sanz, "Cnn-based fisheye image real-time semantic segmentation," in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 1039–1044.

[13] Á. Sáez, L. M. Bergasa, E. López-Guillén, E. Romera, M. Tradacete, C. Gómez-Huélamo, and J. del Egido, "Real-time semantic segmentation for fisheye urban driving images based on erfnet," *Sensors*, vol. 19, no. 1, p. 503, 2019.

[14] T. Sämann, K. Amende, S. Milz, C. Witt, M. Simon, and J. Petzold, "Efficient semantic segmentation for visual birds-eye view interpretation," in *International Conference on Intelligent Autonomous Systems*. Springer, 2018, pp. 679–688.

[15] Y. Wu, T. Yang, J. Zhao, L. Guan, and W. Jiang, "Vh-hfcn based parking slot and lane markings segmentation on panoramic surround view," in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 1767–1772.

[16] X. Zhou, J. Bai, C. Wang, X. Hou, and K. Wang, "Comparison of two panoramic front unit arrangements in design of a super wide angle panoramic annular lens," *Applied optics*, vol. 55, no. 12, pp. 3219–3225, 2016.

[17] 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*, vol. 19, no. 1, pp. 263–272, 2018.

[18] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 2017, pp. 6230–6239.

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

[20] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2015, pp. 1520–1528.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[22] 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*, vol. 39, no. 12, pp. 2481–2495, 2017.

[23] T. Pohlen, A. Hermans, M. Mathias, and B. Leibe, "Full-resolution residual networks for semantic segmentation in street scenes," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 3309–3318.

[24] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 2017, pp. 5168–5177.

[25] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[26] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "Denseaspp for semantic segmentation in street scenes," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 3684–3692.

[27] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid attention network for semantic segmentation," *arXiv preprint arXiv:1805.10180*, 2018.

[28] Y. Yuan and J. Wang, "Ocnet: Object context network for scene parsing," *arXiv preprint arXiv:1809.00916*, 2018.

[29] X. Hu, K. Yang, L. Fei, and K. Wang, "Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation," *arXiv preprint arXiv:1905.10089*, 2019.

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

[31] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context encoding for semantic segmentation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 7151–7160.

[32] 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.

[33] A. Chaurasia and E. Culurciello, "Linknet: Exploiting encoder representations for efficient semantic segmentation," in *Visual Communications and Image Processing (VCIP), 2017 IEEE*. IEEE, 2017, pp. 1–4.

[34] M. Treml, J. Arjona-Medina, T. Unterthiner, R. Durgesh, F. Friedmann, P. Schuberth, A. Mayr, M. Heusel, M. Hofmarcher, M. Widrich *et al.*, "Speeding up semantic segmentation for autonomous driving," in *ML-ITS, NIPS Workshop*, 2016.

[35] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "Icnet for real-time semantic segmentation on high-resolution images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 405–420.

[36] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, and H. Hajishirzi, "Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 552–568.

[37] S.-Y. Lo, H.-M. Hang, S.-W. Chan, and J.-J. Lin, "Efficient dense modules of asymmetric convolution for real-time semantic segmentation," *arXiv preprint arXiv:1809.06323*, 2018.

[38] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 325–341.

[39] T. Wu, S. Tang, R. Zhang, and Y. Zhang, "Cgnet: A light-weight context guided network for semantic segmentation," *arXiv preprint arXiv:1811.08201*, 2018.

[40] X. Chen, X. Lou, L. Bai, and J. Han, "Residual pyramid learning for single-shot semantic segmentation," *arXiv preprint arXiv:1903.09746*, 2019.

[41] K. Yang, L. M. Bergasa, E. Romera, X. Huang, and K. Wang, "Predicting polarization beyond semantics for wearable robotics," in *2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2018, pp. 96–103.

[42] G. Payen de La Garanderie, A. Atapour Abarghouei, and T. P. Breckon, "Eliminating the blind spot: Adapting 3d object detection and monocular depth estimation to 360 panoramic imagery," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 789–807.

[43] S. Siva and H. Zhang, "Omnidirectional multisensory perception fusion for long-term place recognition," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 5175–5181.

[44] R. Cheng, K. Wang, S. Lin, W. Hu, K. Yang, X. Huang, H. Li, D. Sun, and J. Bai, "Panoramic annular localizer: Tackling the variation challenges of outdoor localization using panoramic annular images and active deep descriptors," *arXiv preprint arXiv:1905.05425*, 2019.

[45] Y. Chen, M. Yang, C. Wang, and B. Wang, "3d semantic modelling with label correction for extensive outdoor scene," in *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2019, pp. 1134–1139.

[46] B. Pan, J. Sun, A. Andonian, A. Oliva, and B. Zhou, "Cross-view semantic segmentation for sensing surroundings," *arXiv preprint arXiv:1906.03560*, 2019.

[47] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, "Segmentation and recognition using structure from motion point clouds," in *European conference on computer vision*. Springer, 2008, pp. 44–57.

[48] F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving video database with scalable annotation tooling," *arXiv preprint arXiv:1805.04687*, 2018.

[49] G. Varma, A. Subramanian, A. Namboodiri, M. Chandraker, and C. Jawahar, "Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 1743–1751.

[50] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang, "The apolloscape dataset for autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 954–960.

[51] O. Zendel, K. Honauer, M. Murschitz, D. Steininger, and G. Fernandez Dominguez, "Wilddash-creating hazard-aware benchmarks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 402–416.

[52] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic nighttime image segmentation with synthetic stylized data, gradual adaptation and uncertainty-aware evaluation," *arXiv preprint arXiv:1901.05946*, 2019.

[53] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*. IEEE, 2016, pp. 3234–3243.

[54] S. Wang, M. Bai, G. Mattyus, H. Chu, W. Luo, B. Yang, J. Liang, J. Cheverie, S. Fidler, and R. Urtasun, "Torontocity: Seeing the world with a million eyes," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 3028–3036.

[55] S. Yogamani, C. Hughes, J. Horgan, G. Sistu, P. Varley, D. O'Dea, M. Uricar, S. Milz, M. Simon, K. Amende *et al.*, "Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving," *arXiv preprint arXiv:1905.01489*, 2019.

[56] E. Romera, L. M. Bergasa, J. M. Alvarez, and M. Trivedi, "Train here, deploy there: Robust segmentation in unseen domains," in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 1828–1833.

[57] K. Yang, L. M. Bergasa, E. Romera, and K. Wang, "Robustifying semantic cognition of traversability across wearable rgb-depth cameras," *Applied optics*, vol. 58, no. 12, pp. 3141–3155, 2019.

[58] E. Romera, L. M. Bergasa, K. Yang, J. M. Alvarez, and R. Barea, "Bridging the day and night domain gap for semantic segmentation," in *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2019, pp. 1184–1190.

[59] L. Sun, K. Wang, K. Yang, and K. Xiang, "See clearer at night: Towards robust nighttime semantic segmentation through day-night image conversion," *arXiv preprint arXiv:1908.05868*, 2019.

[60] J. Muñoz-Bulnes, C. Fernandez, I. Parra, D. Fernández-Llorca, and M. A. Sotelo, "Deep fully convolutional networks with random data augmentation for enhanced generalization in road detection," in *Intelligent Transportation Systems (ITSC), 2017 IEEE 20th International Conference on*. IEEE, 2017, pp. 366–371.

[61] S. Liu, J. Zhang, Y. Chen, Y. Liu, Z. Qin, and T. Wan, "Pixel level data augmentation for semantic image segmentation using generative adversarial networks," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 1902–1906.

[62] G. Blott, M. Takami, and C. Heipke, "Semantic segmentation of fisheye images," in *European Conference on Computer Vision*. Springer, 2018, pp. 181–196.

[63] D. Scaramuzza, A. Martinelli, and R. Siegwart, "A toolbox for easily calibrating omnidirectional cameras," in *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*. IEEE, 2006, pp. 5695–5701.

[64] C. Xie, Y. Wu, L. van der Maaten, A. L. Yuille, and K. He, "Feature denoising for improving adversarial robustness," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 501–509.

[65] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 7132–7141.

[66] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 2242–2251.

[67] K. Yang, R. Cheng, L. M. Bergasa, E. Romera, K. Wang, and N. Long, "Intersection perception through real-time semantic segmentation to assist navigation of visually impaired pedestrians," in *2018 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, 2018, pp. 1034–1039.

[68] K. Yang, L. M. Bergasa, E. Romera, D. Sun, K. Wang, and R. Barea, "Semantic perception of curbs beyond traversability for real-world navigation assistance systems," in *2018 IEEE International Conference on Vehicular Electronics and Safety (ICVES)*. IEEE, 2018, pp. 1–7.

[69] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 5122–5130.

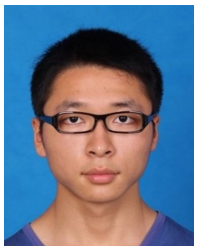
[70] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[71] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 2999–3007.

[72] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[73] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pp. 770–778.

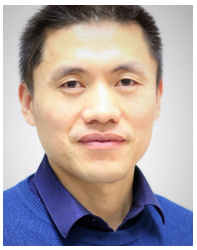
**Kailun Yang** received the B.S. degree in Measurement Technology and Instrument from Beijing Institute of Technology and the dual degree in Economics from Peking University in June 2014. He received the Ph.D. degree in Instrument and Meter Engineering from State Key Laboratory of Modern Optical Instrumentation, Zhejiang University in June 2019. He has also done a research internship in the RobeSafe group at University of Alcalá. His research interests include computer vision, optical sensing, intelligent vehicles, robotics and assistive technologies. Specifically, his Ph.D. project is focused on terrain sensing for navigation assistance systems. For more information, visit his Website: http://www.yangkailun.com/.



**Xinxin Hu** received the B.S. degree in College of Optical Science and Engineer of Zhejiang University in 2017. He is studying for a master's degree in Instrument and Meter Engineering, at the State Key Laboratory of Modern Optical Instrumentation, Zhejiang University. His research interests include optical detection, visual sensing, deep learning, computer vision, 3D vision, semantic segmentation, object detection, knowledge distillation and indoor navigation. Specifically, his master's project is focused on RGB-D image semantic segmentation system for assisted indoor navigation of visually impaired people.

**Luis M. Bergasa** received the MS degree in Electrical Engineering in 1995 from the Technical University of Madrid and the PhD degree in Electrical Engineering in 1999 from the University of Alcalá (UAH), Spain. He is Full Professor at the Department of Electronics of the UAH since 2011. He is author of more than 230 refereed papers in journals and international conferences, and corresponding author of 7 national patents and 1 PCT patent. He was Research Visitor at the Computer Vision Research Group of the Trinity College in Dublin (Irland) in 1998, Visiting Scholar at the Toyota Technological Institute at Chicago (USA) in 2013, and at the OPTIMAL Center Northwestern Polytechnic University (China) in 2017. His research interests include driver behaviors and scene understanding using Computer Vision and Deep Learning Techniques for autonomous vehicles applications. For more information, visit his Website: http://www.robesafe.com/personal/bergasa/.

**Eduardo Romera** received the B.S + M.S. in Telecommunications Engineering and the M.S. in Electronics from the University of Alcalá (UAH), Spain, in 2014 and 2015 respectively. His first M.S. thesis was fulfilled as Erasmus at the Karlsruhe Institute of Technology (KIT), Germany, while he also worked as a research assistant at the research centre Fraunhofer IOSB. He received the Ph.D. degree in Computer Vision and Deep Learning for Autonomous Vehicles at the University of Alcalá (UAH). He has also performed research internships at NICTA/CSIRO (Australia) and UCSD (USA). He is currently the Chief Technology Officier (CTO) at Pixcellence Inc., specialized in AI solutions applied to low-cost and low-power hardware, and a researcher in Artificial Intelligence at the University of Alcalá (UAH), specially focused in Deep Learning applied to perception for self-driving cars. For more information, visit his Website: http://www.robesafe.com/personal/eduardo.romera/.



**Kaiwei Wang** is currently the Deputy Director of National Optical Instrument Engineering Research Center at Zhejiang University. He received a B.S. degree in 2001 and a Ph.D. degree in 2005 respectively, both from Tsinghua University. In October 2005, he started his postdoctoral research at the Center of Precision Technologies (CPT) of Huddersfield University, funded by the Royal Society International Visiting Postdoctoral Fellowship and the British Engineering Physics Council. He joined Zhejiang University in February 2009 and has been mainly researching on intelligent optical sensing technology and visual assisting technology for the visually impaired. Up to date, he owns 60 patents and has published more than 140 refereed research papers. For more information, visit his Website: http://wangkaiwei.org/.