Can we PASS beyond the Field of View? Panoramic Annular Semantic Segmentation for Real-World Surrounding Perception

Kailun Yang[1], Xinxin Hu[1], Luis M. Bergasa[2], Eduardo Romera[2], Xiao Huang[3], Dongming Sun[4] and Kaiwei Wang[1]

***Abstract*— Pixel-wise semantic segmentation unifies distinct scene perception tasks in a coherent way, and has catalyzed notable progress in autonomous and assisted navigation, where a whole surrounding perception is vital. However, current mainstream semantic segmenters are normally benchmarked against datasets with narrow Field of View (FoV), and most vision-based navigation systems use only a forward-view camera. In this paper, we propose a Panoramic Annular Semantic Segmentation (PASS) framework to perceive the entire 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 effort entailed to create 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 smooth and seamless panoramic scene parsing. Beyond the enlarged FoV, we extend focal length-related and style transfer-based data augmentations, to robustify the semantic segmenter against distortions and blurs in panoramic imagery. A comprehensive variety of experiments demonstrates the qualified robustness of our proposal for real-world surrounding understanding.**

I. INTRODUCTION

In the context of intelligent transportation systems, pixel-wise semantic segmentation has attracted rising attention to enable a unified view of semantic scene understanding, universally desired by Intelligent Vehicles (IV) navigation systems [1]. However, almost all semantic perception frameworks are designed to work with conventional sensors capturing a limited Field of View (FoV), such as forward-facing cameras integrated in autonomous vehicles [2] or wearable robotics [3]. Besides, mainstream semantic segmenters are normally benchmarked against conventional FoV images in public datasets, *e.g.*, Cityscapes [4] and Mapillary Vistas [5].

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.

[1]Kailun Yang, Xinxin Hu and Kaiwei Wang are with State Key Laboratory of Modern Optical Instrumentation, Zhejiang University, Hangzhou, China {elnino, hxx_zju, 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

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

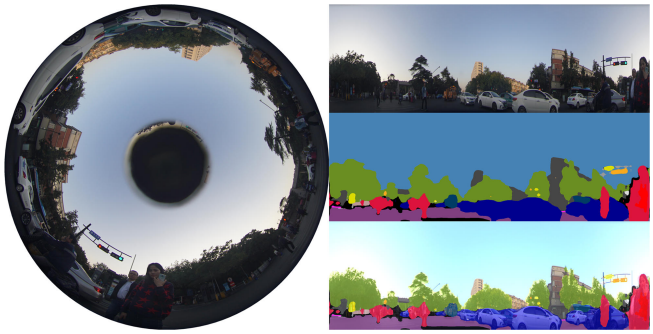


Fig. 1. The proposed panoramic semantic segmentation on real-world surround image captured by our panoramic annular lens system.

This renders semantic segmentation as an insufficient solution to scene understanding, because autonomous/assisted IV need measurably reliable and comprehensive perception of the whole surrounding in order to support upper-level navigational tasks. In this sense, extending semantic segmentation to panoramic perspective is crucial for safe navigation, especially in complex urban scenes such as overcrowded intersections and roundabouts.

To this aim, there are some semantic perception frameworks that have addressed panoramic segmentation by arraying several conventional cameras [2][6][7] or attaching fish-eye cameras with pronounced lens-introduced distortions [8][9][10]. However, these platforms typically require stitching segmented maps from multiple cameras with varying orientations [6][7][8], or only cover less than 180° semantic understanding of the forward surroundings [2], while perception system using only a single camera to achieve 360° panoramic semantic segmentation is scarce in the state of the art. To fill this gap, we propose a Panoramic Annular Semantic Segmentation (PASS) framework using our previously designed Panoramic Annular Lens (PAL) [11], whose compactness is a certainly desirable trait for IV or wearable navigation assistance applications [1]. Another tractable aspect is that its distortion is well maintained in less than 1% and the imaging model follows a clear f-theta law, convincingly appealing for deploying a panoramic semantic perception approach characterized with comprehensiveness.

On the other hand, the fundamental challenge to accomplish this goal lies in the preparation of extensive pixel-accurate annotations which is extremely labor-intensive and time-consuming. Instead, if we could exploit conventional FoV images for training a pixel-wise panoramic segmenter, this would be immensely beneficial for our omnidirecitonal sensing system to support a comprehensive variety of driv-

ing/navigating conditions. To mitigate the data insufficiency, we leverage the public Vistas dataset [5] to yield PASS models bypassing the effort needed to create dense pixel-exact annotations. To preserve the contextual priors in the panoramic content after image unfolding, we adapt our efficient ERF-PSPNet [1][3] to infer semantically different feature maps and fuse them to complete the panoramic segmentation (see Fig. 1) through the last fully convolutional layers. To improve the robustness of PASS model, we apply an extended set of data augmentation methods, earning specialized knowledge in panoramic content towards real-world autonomous/assisted navigation. To the best of our knowledge, this is the first panoramic semantic segmentation framework using a PAL system without stitching segmented maps from multiple cameras nor unwrapping fish-eye images with remarkable distortions.

We have already presented some preliminary studies on pixel-wise semantic segmentation combined with unified terrain awareness [1] and pixel-wise polarization prediction [3]. In this paper, we explore on the comprehensiveness aspect where the central contribution is the proposed PASS framework. Additionally, the following bullets sum up technical innovations delivered in this paper:

- A PASS pipeline involving annular image unfolding, semantic feature map predicting and fusing to fulfill pixel-wise segmentation in panoramic imagery. Code associated with replication of the experiments and implementations will be open-sourced at[1].
- A wearable prototype with an omnidirectional PAL system for capturing and collecting panoramic images.
- A finely annotated semantic segmentation dataset for benchmarking panoramic perception algorithms and evaluating real-world performance. The PASS dataset is publicly offered to the community.
- A cluster of specialized ring-padding, cross-segment padding, and upsampling operations to enable true 360° semantic scene parsing.
- An extended set of focal length-related and style transfer-based data augmentations to attain robustness against distortions and blurs in panoramic content.

II. RELATED WORK

A. Fish-eye and Panoramic Semantic Segmentation

Omnidirectional vision sensors are capable of capturing larger FoV of surrounding scenes than traditional pinhole cameras, whose proliferation is accelerating for being mounted on retrofitted IV perception platforms. However, contemporary works using omnidirectional cameras have predominantly focused on visual localization [12] and monocular depth estimation [13]. In contrast, panoramic semantic segmentation, which has not been explicitly investigated, should be traced back to fish-eye image parsing. L. Deng *et al.* [9] overlapped pyramidal pooling [14] of encoded feature maps for fish-eye image segmentation that theoretically facilitates the entire understanding of frontal hemispheric view. They extended the work by using four wide-angle cameras to build a surrounding view system, restricted the deformable convolution to learn geometric transformation and keep spatial structure within fish-eye perspective [8]. Á. Sáez *et al.* [10] followed this trend by implementing real-time fish-eye image segmentation and outperformed the seminal work [9] in terms of both inference speed and accuracy using ERFNet [15], where the eventual goal was to complement a LiDAR sensor installed in their autonomous vehicle.

W. Zhou *et al.* [2] replaced a 56° FoV camera with three 100° FoV lens in an array, aimed to parse a full forward-facing panorama by stitching the undistorted fish-eye segmentation maps. However, it was still only able of perceiving the surroundings in front of the vehicle. To enable panoramic automotive sensing in urban environments, R. Varga *el al.* [6] proposed a super-sensor with four fish-eye cameras, whose images were segmented by using a boosted forest and unwarpped 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, pursuing perception wideness around a test vehicle, five cameras were placed equiangularly on top of the instrumented car by K. Narioka *et al.* [7]. They trained only with front-facing camera images to maintain compactness of Convoluntional Neural Networks (CNNs) and grasped all-direction semantics and depth estimations. Unlike these methods that need to be modeled in complex separate ways, we aim to use a single camera to comprehensively parse real-world scenes and revive the PAL-based vision research line.

B. Semantic Segmentation Datasets and Data Augmentations

Propelled by the breakthrough of deep learning, semantic segmentation has approached the crucial point to unify diverse perception tasks desired by autonomous/assisted navigation systems. While encouraging, semantic segmentation based on CNN architectures has evolved a high reliance on the visual data, from which semantically meaningful features are directly learned in the training process. Naturally, semantic segmentation datasets have played an essential role and spurred key creativity in IV research field. In the last years, numerous IV-oriented datasets have emerged such as Cityscapes [4] and Mapillary Vistas [5]. Cityscapes is one of the milestone benchmarks with videos taken from a camera behind the windshield of the IV, which only offers forward-view of semantic urban scene understanding. Vistas is a large-scale segmentation dataset that attains global geographic reach of observations from different continents, and extends front-facing perspective to diverse capturing viewpoints (*e.g.*, from roadways, sidewalks, unconstrained environments and off-road views), with promising implications in a broad variety of robotic vision applications [3]. These datasets have thousands of images, but even their variety does not assure satisfactory results of current segmenters in unfamiliar domains. Synthia [16] was proposed to facilitate learning with synthetic images. Its virtual ac-

[1]Code and Dataset: `github.com/elnino9ykl/PASS`

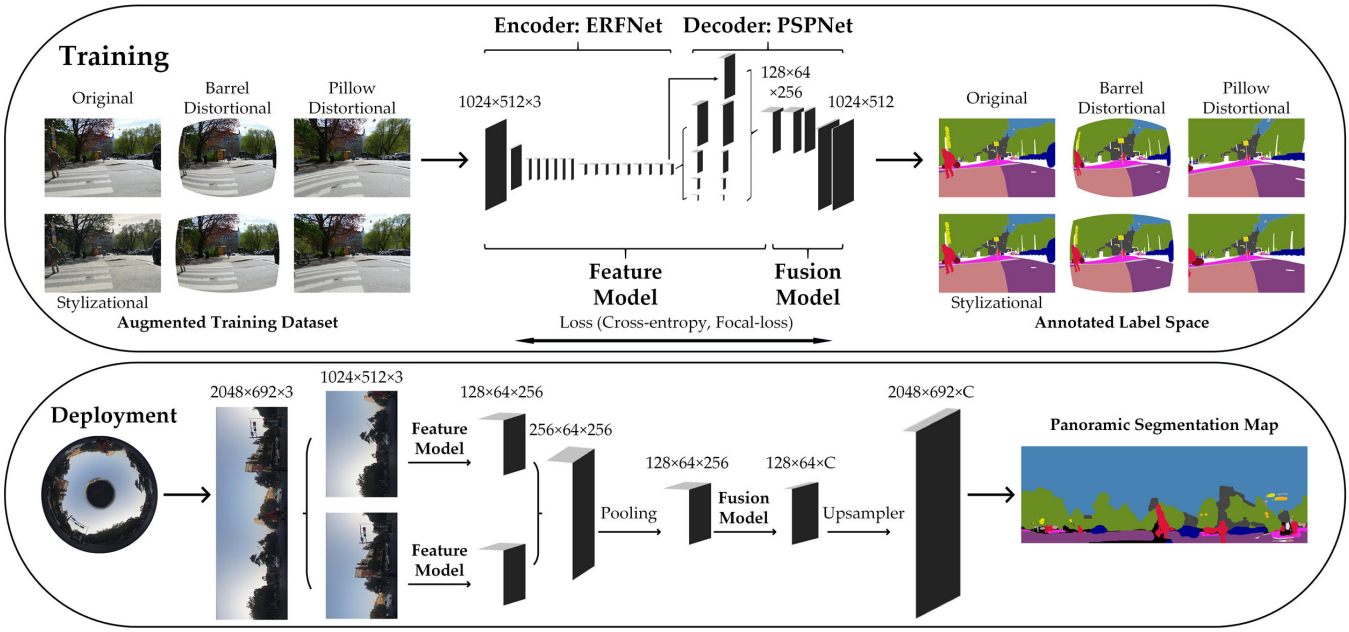


Fig. 2. The proposed panoramic annular semantic segmentation framework.

quisition platform incorporates four 100° FoV binocular cameras with certain overlapping that can be used to create an omnidirectional view as validated in [8]. The TorontoCity benchmark [17] collected spherical panoramas from both drones and vehicles with pixel-level roadway annotations. While these outdoor panoramic datasets are serviceable, significant distortions were introduced which are not compatible with images captured by our PAL system.

Under the vital topic of robust scene understanding, data augmentations have been broadly adopted to expand the datasets and combat overfitting as an implicit regularization technique. To robustify against unseen domains, a recent effort has been made to separate augmentations between geometry and texture, which also achieved desired network calibration [18]. Regarding specialized augmentation methods to earn robustness against distortions that are inevitable when cameras have very large FoV as pointed out in [6], skew and gamma corrections were investigated in [2], while zooming policy was designed for fish-eye images in [9]. In this work, we extend the zoom alteration and combine it with style transfer-based augmentation for panoramic semantic segmentation, and perform a systematic analysis on the robustness gains at the deployment level when trained with conventional FoV images, despite the ubiquitous use of data augmentation, to elevate accuracy on benchmark databases or real navigation systems.

III. Framework

A. Training Stage

The overview of the proposed PASS framework is depicted in Fig. 2. In the training stage, our publicly available semantic segmentation network ERF-PSPNet [1][3] is adapted, which is built using an efficient encoder from ERFNet [15] and a pyramidal pooling-based decoder from PSPNet [14]. Our ERF-PSPNet inherits both technical gists including spatial factorized filters, sequential/hierarchical dilations and pyramid representations, so as to strike an essential balance between real-time speed and accuracy, and outperforms ERFNet in context-critical domains [1] while maintaining the compactness to be easily deployed in an embedded system such as NVIDIA Jetson TX1/TX2[2]. By training on a conventional FoV imaging dataset, an efficient 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 accurately predicted by F that can also be separated into a feature model F_e and a fusion model F_u, formally:

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

In this work, we re-purpose it to address 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 image is unfolded using the interface provided by the omnidirectional camera toolbox [19]. The unfolded panoramic image is partitioned into M segments as it is depicted in the following equation:

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

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, the straightforward solution is to directly integrate the inferred pixel-wise probability maps

[2]ERF-PSPNet: `github.com/dongmingsun/tx2-erfpsp`

of M segments along the unfolding direction. 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 panoramic segments and features inferred from the conventional images used in the training:

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

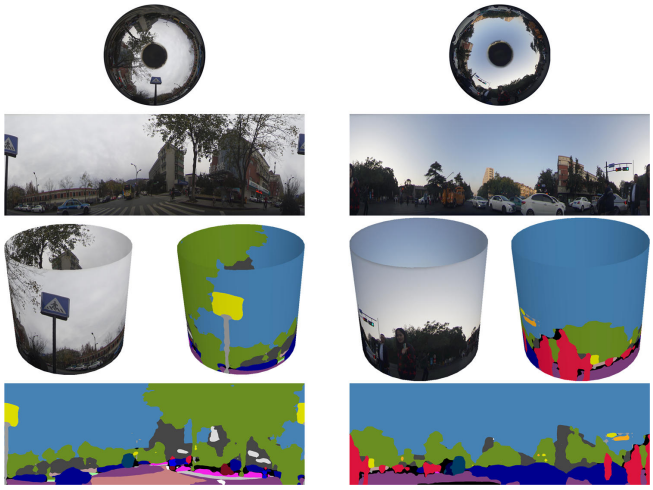
After the concatenation of the M segments and a max-pooling process to recover the original feature model size, the entire panoramic annular image is smoothly parsed by the fusion model F_u, since semantically abstract features have already been extracted and aggregated. Followed by a bilinear upsampler, the final panorama segmentation map $S_p^{H_p \times W_p}$ is obtained by matching to the inputting size:

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

where $\biguplus$ denotes concatenation of feature maps, which can also be considered as a feature denoising block to increase robustness when added before the 1×1 convolution layer [20]. For illustrative purposes, M is set to 2 in Fig. 2. In our experiment section, different settings (M =1, 2, ..., 6) are explored and compared to study the effect of this FoV-related parameter on the final 360° segmentation performance.

C. Network Adaptation

We propose some network adaptation techniques to face borders discontinuity in the panorama (see Fig. 3) or overlapping of the 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 left and right borders together as depicted in Fig. 3. This operation was first introduced as ring-padding in [13] 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 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]. Accordingly, the padding has been proportionally widened to the dilation rate. We also 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. Additionally in the bilinear interpolation layers of our decoder, we include specialized ring-upsampling and cross-segment upsampling to eliminate the undesirable boundary effects.



(a) (b)

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

D. Data Augmentation

Our purpose is to learn from conventional 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 augmentations, random rotation and shear are firstly implemented with degrees both uniformly sampled from the angles [-1°, 1°] to change positions of pixels while maintaining lines straight. Followed by rotation/shear transformation, we implement translation and aspect-ratio augmentation. 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 is individually performed at a 50% opportunity to improve orientation invariance. Regarding textural changes, brightness, contrast, saturation and hue variations are simultaneously augmented by selecting the values in random within the range [-0.1, 0.1] to improve the robustness against diverse illumination conditions and 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 [9][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 imaging point $P_c = (x_c, y_c)$ by adjusting the distance to the principal point P_p:

$$r_c = f \times tan(r_a/f),$$

where r_c denotes the distance between the point P_c and the principal point P_p on the conventional image, while r_a correspondingly denotes the distance between the P_a the principal point P_p on the augmented image. This mapping helps to add robustness against barrel distortion that is common in fish-eye images. In this work, we extend the augmentation to address both barrel and pillow distortions by additionally creating training samples with adjusted distance:

$$r_c = f \times arctan(r_a/f).$$

Although this set of distortional augmentations doesn't strictly follow the PAL imaging law, the joint use with geometric and textural 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 [11]. 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-quality VISTAS imagery, and a critical part of panoramic images are captured in hazy weather and low illuminated conditions. To improve the robustness of semantic segmenters when taken outside their comfort zones, we leverage CycleGAN [21] to learn a transformation back and forth between the VISTAS and our PASS that are two unpaired domains. We incorporate transformed training images from Vistas while preserving the original geometry of semantic labels as additional samples. In this way, the GAN-based transfer is used as a stylizational data augmentation technique to robustify against the blurs and compression artifacts present in panoramic imagery. Otherwise, the lack of invariance to blurring may bias the segmenter and corrupt the prediction when learning from total high-quality images.

IV. Experiments and Discussion

A. Experimental Setup

The segmentation performance is evaluated on the Mapillary VISTAS validation dataset and our Panoramic testing dataset (PASS dataset), which is collected by building a personal navigation assistance device with the compact PAL system that captures a FoV of 360°×75° (30°-105°), as worn by the user in Fig. 4a. The PASS dataset contains 1050 raw and unfolded panoramic annular image pairs, from which 400 panoramas are finely labeled with masks on 4 of the most critical classes for navigation assistance: *Car*, *Road*, *Crosswalk* and *Curb*. Schematically, four unfolded ground-truth annotated images are shown in Fig. 4b.

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

(a) (b)

Fig. 4. (a) Wearable navigation assistance system; (b) Unfolded panoramas with annotations.

numerical results are gathered by using the prevailing "Intersection over Union" ($IoU = \frac{TP}{TP+FP+FN}$) or "Global Accuracy" ($ACC = \frac{CCP}{LP}$) for fair comparisons.

B. Training Setup

The Mapillary Vistas dataset [5] is utilized for training our semantic segmentation models to take advantage of its wide coverage and high variability in observation viewpoints, other than learning with only forward-view images [4][7]. 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 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 capacities to densely interpret metropolitan scenes. Regarding the CNN training setup, we train all models under the same conditions using Adam optimization [22] with an original Learning Rate (LR) of 5×10^{-4} and Weight Decay (WD) of 2×10^{-4}, exponentially decreasing LR until the loss converges when feeding images at batch size of 6 and resolution of 1024×512 on a single GPU NVIDIA 1080Ti. The 2D version of focal loss [3] is adopted as the training criterion instead of conventional cross entropy. Under this setup, our ERF-PSPNet reaches mean IoU (mIoU) of 54.3% on the validation dataset. This result achieved without any data augmentation is marked as the baseline, where per-class accuracy values are displayed in Table I(a), 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 critical to investigate the influence of the number of segments (M) on the final performance in panoramic view. In this experiment, the unfolded panoramic image has been partitioned into 1, 2, ..., 6 segments, corresponding to a FoV of 360°,

TABLE I

ACCURACY ANALYSIS.

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) On the Accuracy of Semantic Segmentation. ***Pol***, ***StL***, ***Bil*** etc. are abbreviations of the classes.

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%**

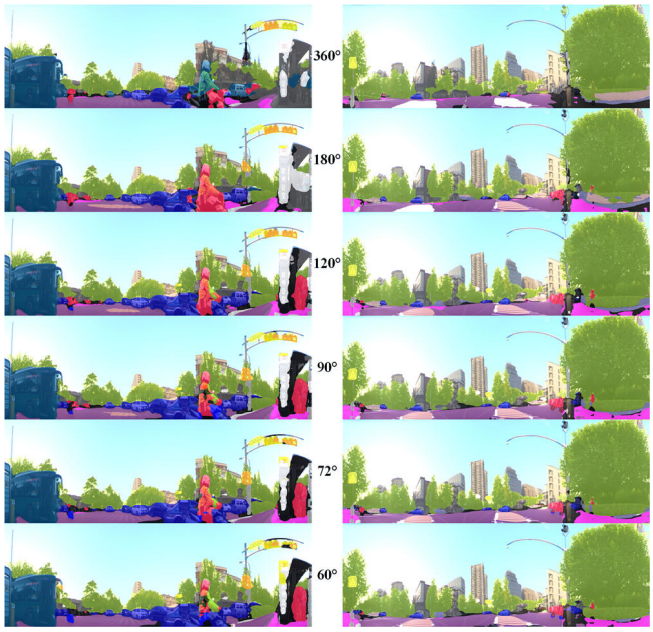
(b) On the Influence of Number of Segments. **Blue** denotes higher IoU with specialized padding and upsampling. **Red** highlights the best IoU.

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%**

(c) On the Robustness of Panoramic Segmentation.

180°, ..., 60° per segment. As displayed in Table I(b), if you use only a feature model for the whole panorama, the context is too wide and results are worse than when the segment is more adapted to exploit the features of the classes. Consequently, the 360°-per-segment model suffers from a large loss of accuracy with the incompatible contextual cues. In comparison, the 180°-per-segment predictor achieves the highest IoU on roadway segmentation, while the 120°-per-segment predictor exceeds other solutions 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 case. Vitally, regardless of segments number, the use of specialized padding and upsampling helps to boost the accuracy in almost all conditions, demonstrating the effectiveness of our network adaptation proposal as one of the key enablers to fulfill 360° semantic segmentation.

The segments finding is also consistent with the qualitative results. As comparably visualized in Fig. 5, the 360°-per-segment results are undesirable with limited detectable range of traversable areas, *e.g.*, roadways and sidewalks. In Fig. 5a, the 360°-per-segment model wrongly colors the rider, and 360°/180°-per-segment models cannot detect the person on the right side. Intriguingly in Fig. 5b, the 360°-per-segment approach has classified both crosswalk areas as general road markings, while the 180°-per-segment solution only correctly identifies a crosswalk region. Our plausible hypothesis is that in most of the training samples, only one crosswalk region will be observed, hence 120°/90°-per-segment models are better at crosswalk detection as well as the segmentation of diverse vehicles and curbs. On the other side, when using more feature models ($M = 5$ or 6), the segmentation tends to become fragmented. To maintain a good trade-off, we set M to 4 in the following experiments.



(a) (b)

Fig. 5. 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.

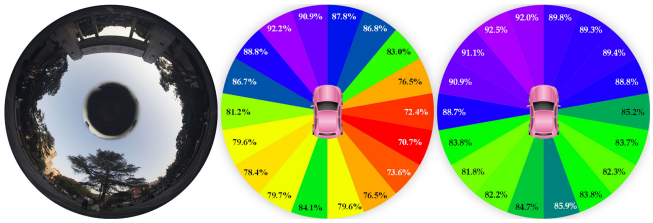
D. On the Robustness of Panoramic Segmentation

Taking an essential stride to delve into "accuracy" and "robustness", the gap between these two concepts can be better understood in the context of panoramic semantic segmentation. We collect the segmentation accuracy on the densely annotated Vistas validation dataset, in contrast with the real-world accuracy on PASS for testing (both in IoU), as displayed in Table I(c). 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 significantly boosted in terms of cars and roadways that 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 our PASS dataset are not as illuminated as most scenarios from Vistas, needless to mention that the augmented aspect ratio is critical for panoramic segmentation. Traditional augmentation also implies a slight accuracy decrease on validation dataset as the accuracy in the unseen panoramic domain greatly increases, which further gives an intuition on how augmenting data highly prevents overfitting and helps yielding robust models for deployment. 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 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 remarkably improved, which is due to that within panoramic imagery, most of the crosswalks are not as clear as those in Vistas dataset. The style transfer algorithm excels exactly at generating realistic blurs with an example visualized in Fig. 2, thus making the augmented model more prepared against the panoramic domain. When combining all heterogeneous data augmentations, the best accuracy boosts have been achieved for all classes, outperforming any independent augmentation by a large margin. 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. 6 also demonstrates the effectiveness of data augmentation, which is a visualization of roadway segmentation accuracy values (ACC) on the panoramic dataset by using the PASS model without/with data augmentation. Following [12], we partition the panoramic image into 18 directions and notice 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 IV and wearable personal guidance systems. Fig. 7 showcases diverse segmentation maps in challenging frames of our PASS dataset. It can be easily seen in all qualitative segmentation examples of both campuses and complex intersections, our augmented model delivers impressive 360° semantic segmentation despite the distortions and blurs, owing to the proposed framework and the extremely positive effect of data augmentation in refining and robustifying panoramic segmentation.



(a) (b) (c)

Fig. 6. (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.

V. Conclusions and Ongoing Work

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 IV or wearable assisted navigation systems with advanced capabilities to accurately interpret the surroundings in a universal and comprehensive manner. While the same panoramic view can be achieved from 4-6 cameras surrounding a vehicle with high resolution which is normally kept as redundancy, our system only uses a single camera. The proposed 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.

Towards long-term navigation assistance, robotic vision characterized with wide FoV and multidimensionality [3], is a key source of momentum in our ongoing project. Since we have identified new challenging automotive sensing problems like the cross-perspective context issue, it is worthwhile contributing future efforts by synthesizing semantics that conform to panoramic imagery. In addition, we will expand the PASS dataset by labeling more classes and design relevance-aware loss functions for safety-critical applications.

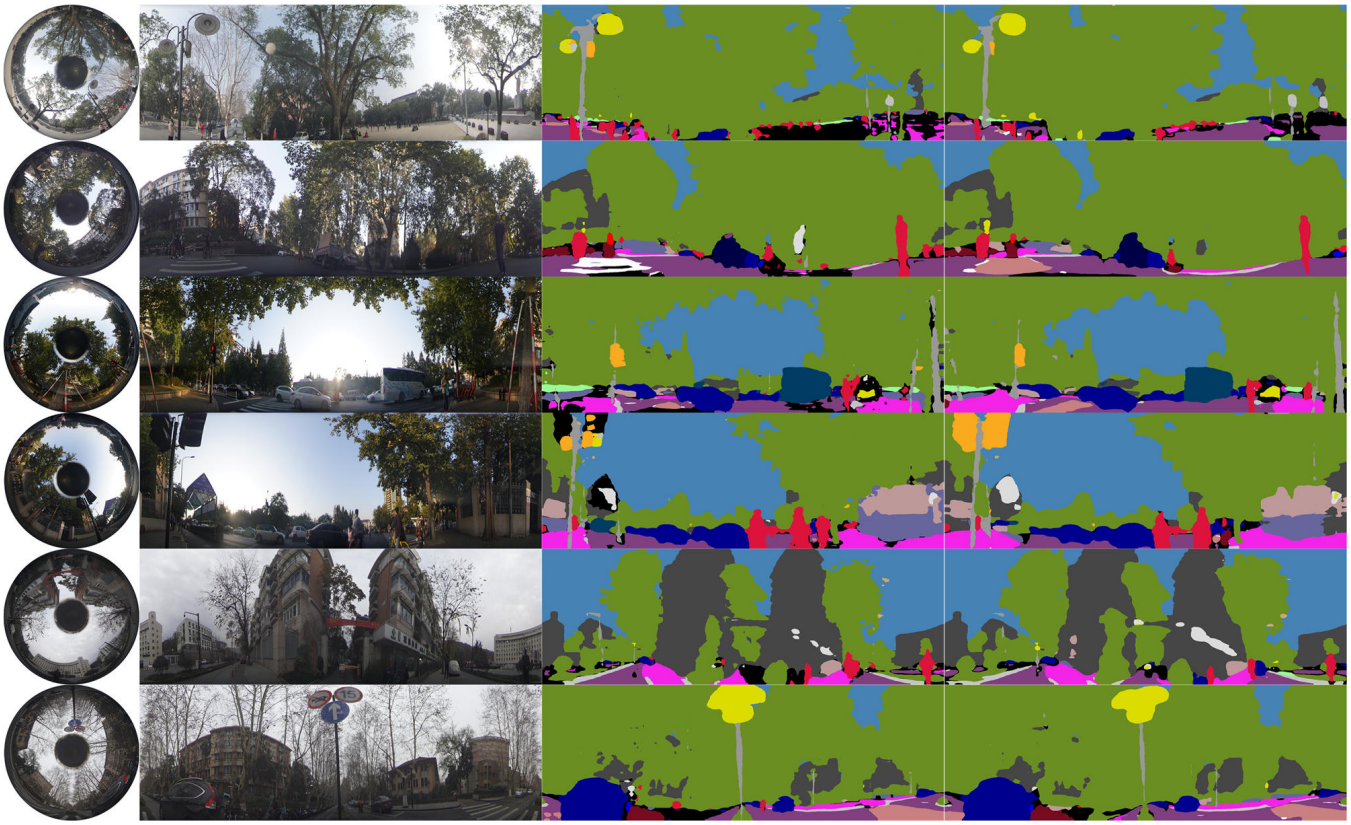
References

[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] 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.

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

[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 *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*. IEEE, 2016, pp. 3213–3223.



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

Fig. 7. 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 data augmentation.

[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] 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.

[7] 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.

[8] 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.

[9] 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.

[10] A. 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.

[11] X. Huang and J. Bai, "Analysis of the imaging performance of panoramic annular lens with conic conformal dome," in *AOPC 2015: Optical Design and Manufacturing Technologies*, vol. 9676. International Society for Optics and Photonics, 2015, p. 96760G.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] C. Xie, Y. Wu, L. van der Maaten, A. Yuille, and K. He, "Feature denoising for improving adversarial robustness," *arXiv preprint arXiv:1812.03411*, 2018.

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

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