Is Context-Aware CNN Ready for the Surroundings? Panoramic Semantic Segmentation in the Wild

Kailun Yang[1], Xinxin Hu[2], and Rainer Stiefelhagen[1]

***Abstract*—Semantic segmentation, unifying most navigational perception tasks at the pixel level has catalyzed striking progress in the field of autonomous transportation. Modern Convolution Neural Networks (CNNs) are able to perform semantic segmentation both efficiently and accurately, particularly owing to their exploitation of wide context information. However, most segmentation CNNs are benchmarked against pinhole images with limited Field of View (FoV). Despite the growing popularity of panoramic cameras to sense the surroundings, semantic segmenters have not been comprehensively evaluated on omnidirectional wide-FoV data, which features rich and distinct contextual information. In this paper, we propose a concurrent horizontal and vertical attention module to leverage width-wise and height-wise contextual priors markedly available in the panoramas. To yield semantic segmenters suitable for wide-FoV images, we present a multi-source omni-supervised learning scheme with panoramic domain covered in the training via data distillation. To facilitate the evaluation of contemporary CNNs in panoramic imagery, we put forward the Wild PAnoramic Semantic Segmentation (WildPASS) dataset, comprising images from all around the globe, as well as adverse and unconstrained scenes, which further reflects perception challenges of navigation applications in the real world. A comprehensive variety of experiments demonstrates that the proposed methods enable our high-efficiency architecture to attain significant accuracy gains, outperforming the state of the art in panoramic imagery domains.**

***Index Terms*—Scene Understanding, Semantic Segmentation, Panoramic Images, Scene Parsing, Autonomous Driving.**

I. Introduction

SEMANTIC segmentation, enabling a unification of navigational perception tasks [1], catalyzes striking progress in autonomous transportation. Convolutional Neural Networks (CNNs) achieve superior performance at this task [2]. In particular, thanks to the capacity to efficiently exploit wide context information, modern CNNs are able to fulfill semantic segmentation both swiftly and accurately [3][4].

However, most of current CNNs are designed for narrow Field of View (FoV) pinhole images in mainstream datasets like Cityscapes [5] and Mapillary Vistas [6], while a 360° semantic segmentation benchmark is rare in the state-of-the-art. Despite the growing popularity of panoramic cameras for a complete sensing of the entire surrounding [4], semantic segmenters have not been comprehensively evaluated on omnidirectional wide-FoV images. Nevertheless, it is highly uncertain whether a context-aware CNN maintains its performance when taken to unseen panoramic scenes with the gap between real-world domains, as well as the disparity of structural information between pinhole images and full-view panoramas, being evident.

To illustrate this issue, Fig. 1 displays examples of panoramic images from our WildPASS database, which is put forward as a new evaluation dataset to kindle the research on surrounding perception. It can be seen that the panoramas feature rich and distinct global contextual cues as various directions of roadways and sidewalks are simultaneously imaged. Fig. 1 further depicts the semantic distribution of the

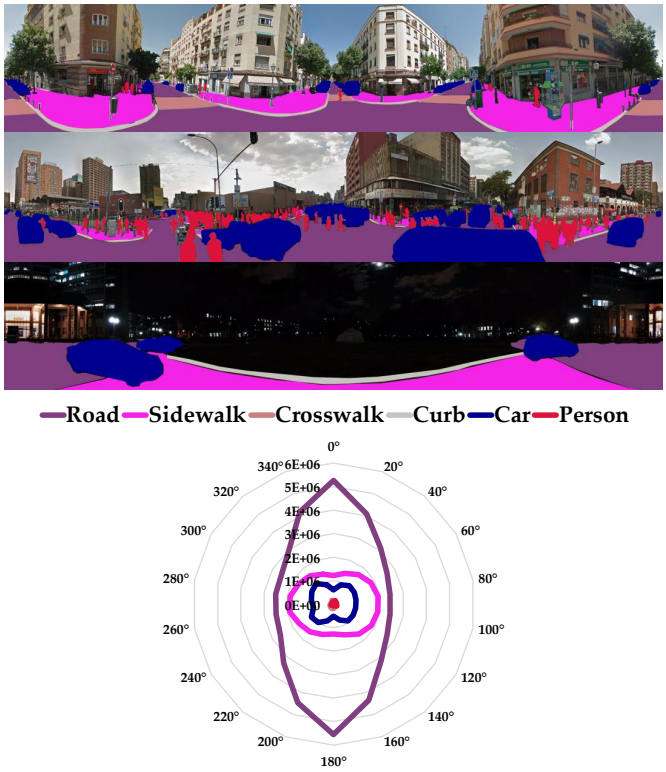


Fig. 1. Top: Challenging panorama examples from our Wild PAnoramic Semantic Segmentation (WildPASS) dataset, including omni-direction roadways, densely populated areas and nighttime scenes; Bottom: Class distributions of the WildPASS dataset in terms of the number of pixels in different directions.

Manuscript received May 23, 2020; revised October 10, 2020; revised November 22, 2020; accepted December 25, 2020. This work was supported in part by the Federal Ministry of Labor and Social Affairs (BMAS) through the AccessibleMaps project under Grant 01KM151112, in part by the University of Excellence through the "KIT Future Fields" project, in part by the KIT-Publication Fund of the Karlsruhe Institute of Technology, in part by the Hangzhou SurImage Technology Company Ltd., and in part by the Hangzhou KrVision Technology Company Ltd. (krvision.cn). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xiaolin Hu. *(Corresponding author: Kailun Yang.)*

[1]Kailun Yang and Rainer Stiefelhagen are with Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany. (e-mail: kailun.yang@kit.edu; rainer.stiefelhagen@kit.edu).

[2]Xinxin Hu is with Huawei Technologies Company Ltd., Hangzhou 310000, China. (e-mail: anheidelonghu@gmail.com).

WildPASS dataset where the road class is centered around both 0°/360° and 180° (front-facing) directions, clearly different to that of pinhole forward-view images [5]. However, the potential of the context information implicated in panoramic scenes is not easily materialized without large-scale training panoramas and explicit distinction of the contextual cues.

Motivated by this observation, this paper proposes a concurrent horizontal and vertical attention module, to jointly and explicitly, extract width-wise and height-wise contextual information markedly available in panoramas. The width-wise contextual information that represents the context along the horizontal direction, as analyzed above, supposes rich priors in ultra-wide FoV images. The height-wise context, orthogonally, encodes structural information in the vertical direction, which are shared as a common nature of street scenes [7]. With the concurrent attention embedded into a deep CNN, the feature maps are recalibrated to be more meaningful along space for panoramic segmentation, both horizontally and vertically.

To yield semantic segmenters suitable for panoramas, we argue that it is essential to expose the learner to wide-FoV omnidirectional inputs before it is ready to be deployed in the wild. We present a multi-source omni-supervised learning scheme to cover panoramic imagery domain in the training of efficient CNNs. Concretely, the omni-supervised learning concept is approached through data distillation [8], where the efficient learner CNN exploits both labeled pinhole images and unlabeled 360° full-view panoramas whose wide-angle and wrap-around connections are considered.

To address the scarcity of panoramic scene parsing testbeds, we put forward the Wild PAnoramic Semantic Segmentation (WildPASS) database with pixel-wise annotations using classes defined in Mapillary Vistas, as a benchmark with new metrics to facilitate credible numerical evaluation and comparison of state-of-the-art semantic segmentation CNNs in panoramic imagery. This dataset embraces the wild by extracting street-view panoramas from all around the world (6 continents, 25 cities), which also includes highly unconstrained environments and adverse conditions such as the nighttime (see Fig. 1). WildPASS offers a high variability of capturing viewpoint, with images taken from both roadways and sidewalks, a more comprehensive reflection of the challenges for real-world autonomous navigation systems to semantically understand their surroundings.

An extensive set of experiments shows that the concurrent attention and the omni-supervised training, both help the efficient context-aware learner to attain significant accuracy boosts and generalization gains in unseen panoramic domains. With these two key enablers, our high-efficiency architecture produces state-of-the-art performances on the public PASS [9] and the fresh WildPASS datasets. In summary, we deliver the following contributions:

- We rethink panoramic image semantic segmentation from the context-aware perspective and propose a concurrent attention module to exploit height-wise and width-wise contextual information in panoramas across the 360°.
- We put forward a multi-source omni-supervised learning scheme to cover panoramic imagery in training and unlock the potential of global contextual priors for omnidirectional perception.
- We present the diverse WildPASS dataset, which is collected from all around the globe for evaluating panoramic segmentation in the wild and will be open-sourced at.[1]
- We largely elevate state-of-the-art single-pass segmentation performances on both PASS and WildPASS datasets.

II. Related Work

A. From Accurate to Efficient Semantic Segmentation

Semantic segmentation has shown tremendous progress since Fully Convolutional Networks (FCNs) [2] that illuminated the vista of end-to-end pixel-wise prediction. SegNet [10] laid the foundation of the encoder-decoder structure. Benefiting from deep backbones in classification architectures like ResNet [11] and DenseNet [12], segmentation CNNs such as DeepLab [13], PSPNet [14] and DenseASPP [3] greatly advanced the performance frontiers on a few public datasets. Pyramid pooling [14], atrous spatial pyramid pooling [3][13] and multi-scale feature representations [15][16][17] helped to sense different levels of surrounding context. SPNet [18] employed strip pooling to aggregate diverse spatial information, together with pyramid pooling for short-range dependency collection. The boundary-aware feature propagation module [19] was proposed to harvest local features within their regions isolated by the learned boundaries.

Recently emerged attention mechanisms [20][21][22] allow to spotlight informative features and thereby have been applied in vision tasks. One of the well-known attention operation introduced in Squeeze-and-Excitation network (SENet) [20], leveraged global average pooling to squeeze spatial information and capture full-view context, followed by an excitation block that re-weights the feature map according to the importance of different channels. In semantic segmentation, such attention means facilitated the exploitation of global statistics (EncNet [23]), cross-modal complementary features (ACNet [24]) and height-wise structural priors (HANet [7]). To capture long-range dependencies, non-local blocks [25] are widely used. DANet [26] took advantage of self-attention operations [27] to aggregate pair-wise relations between any two pixels or channels, while OCNet [28] modeled object context. To reduce the complexity of the pair-wise computations, CCNet [29] adopted a criss-cross module to refine the context along horizontal and vertical directions for each pixel. ORDNet [30] restricted the self-attention into local patches to emphasize middle-range dependencies. In this work, we propose an concurrent attention module, which uses height-wise and width-wise pooling to aggregate spatial information, while spotlighting the important horizontal and vertical positions by boosting the informative features. Our concurrent attention module fundamentally differs from existing context aggregation methods [18][29] as our attention aims to indicate which channels are critical at each individual column or row, instead of only aggregating the spatial dependency or refining the features based on feature similarity. Thereby, our module is promising to stabilize wide-FoV semantic segmentation.

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

The preceding efforts all contribute to the exceedingly high scores on forward-view pinhole image benchmarks. However, the divergence of global contextual information between pinhole and omnidirectional data leads to the open question whether context-aware models are qualified in open panoramic domains. Also, for panoramic images, semantic segmentation is required to perform with megapixel resolution to cover the 360°, which intensifies the computation constraint and disqualifies the usage of top-performance sophisticated architectures in real-time applications. This shifted the research efforts to design efficient architectures. To list a few representative networks, ENet [31], SQNet [32], ERFNet [33], ICNet [16], LinkNet [34], ESPNet [35], EDANet [36], BiSeNet [37], CGNet [38] and SwiftNet [39]. In previous works, we built up our solution with ERF-PSPNet [1] and SwaftNet [4] for autonomous navigation systems. These light-weight networks have struck a better trade-off between accuracy and inference latency thanks to the efficient use of context information. However, most of them were targeted to forward-view clean images with limited FoV where the complexity is also limited to measure the reliability and applicability of segmentation CNNs in the wild.

B. From Surround-View to Panoramic Semantic Segmentation

Early vision-based surrounding perception systems leveraged fisheye sensors [40] that were often horizontally arranged in an array [41][42]. Another cluster of semantic perception platforms [43][44] resorted to multiple cameras to attain the full 360° coverage. Nevertheless, generating on the fly a holistic representation incurs significant latency and computational complexity to process multi-view images, coming with a sequence of fundamental yet hard works such as camera calibration, semantic mapping and data fusion [42]. Also, the size and cost of multiple cameras or blind spots between different lenses could be problematic in some scenarios [9].

Recently, one of the trendiest approaches is to use a single panoramic camera [45][46] by running segmentation CNNs directly on the 360° image, which are usually trained on self-collected data such as the CamVid-360 dataset that was gathered along the same path of the well-known CamVid database [46]. With similar purposes, Omni-SYNTHIA [45] and SYNTHIA-PANO [47] were generated from virtual collections. The purely synthetic OmniScape [48] also belongs to this group. However, the large discrepancy between synthetic and real-world imagery entails further domain adaptation [49][50] which often needs to adapt the inputs on-the-fly that sacrifices efficiency during inference, yet struggles to generalize in open domains. While there are a few wide-angle databases such as TorontoCity [51] and WoodScape [52], their scene variety and annotation density are far lower than pinhole benchmark datasets like Cityscapes [5], Mapillary Vistas [6], IDD [53] and BDD [54], which suppose rich ontologies for producing robust segmenters.

Contrary to those approaches trained with limited data of panoramic scenes, Yang et al. [9][55] presented a Panoramic Annular Semantic Segmentation (PASS) framework, whose underlying paradigm is to train on pinhole data and deploy in unseen panoramas. This unlocks the usage of panoramic cameras in a comprehensive variety of application scenarios by taking advantage of the wealth of large-scale datasets. Unlike previous multi-view systems, they did only need a single process to fulfill continuous panoramic segmentation. However, it was achieved by separating the panorama into several segments, each of which is independently fed to the encoder to obtain feature maps for the final fusion, significantly increasing the running time and computation burden [9]. Besides, with the panorama partitioned into pieces for several forward passes, they miss the opportunity to leverage the crucial width-wise context information available in the full-view scene.

Unlike the previous works, we aim to fulfill omnidirectional semantic segmentation in a single pass and thereby propose a multi-source omni-supervised learning scheme to cover panoramic imagery in the training. The omni-supervised concept is approached through data distillation [8], which is different to conventional model distillation methods [56][57] that ensemble multiple experts but entail re-training different cumbersome networks. Instead, we create a light-weight ensemble of data transformations of panoramic images. Compared to the pseudo-labeling methods [58][59] in video semantic segmentation and omni-supervised learning in human key point detection [8] that require massive amounts of extra sequences, we only use a moderate number of unlabeled panoramas, while significantly stabilizing omnidirectional scene parsing in the wild. Specifically, our omni-supervised learning scheme operates in a multi-source manner to add the diversity of FoV, intertwining both densely labeled labeled pinhole images and unlabeled panoramas, where the wide-angle and wrap-around structures of panoramic images are fully considered. Finally, the learned efficient CNN can perform both swift and accurate omnidirectional image segmentation in the wild without having to manually label any panorama for training.

III. Methodology

Panoramic images, suppose important contextual priors depending on a spatial position, due to the wide FoV to cover horizontal 360° surroundings and the common vertical structures for street scenes. To leverage the rich structural information, we propose a concurrent horizontal and vertical attention module to jointly encode width-wise and height-wise context in a series of steps, which are elaborated in Sec. III-A. To cover the panoramic imagery domain in the training, a multi-source omni-supervised learning scheme is presented, which we describe in detail in Sec. III-B the preparation, training and deployment stages.

A. Concurrent Horizontal and Vertical Attention

Following recent architectures [7][20][21] that use attention mechanism to spotlight useful features, we aim to design a module that allows to indicate important features at individual columns and rows, and thereby stabilize the segmentation of wide-FoV street-view images. Specifically, inspired by the SENet [20] that calibrates feature maps by squeezing spatially and exciting channel-wise, we propose a concurrent attention module, which explores alternative dimensions of

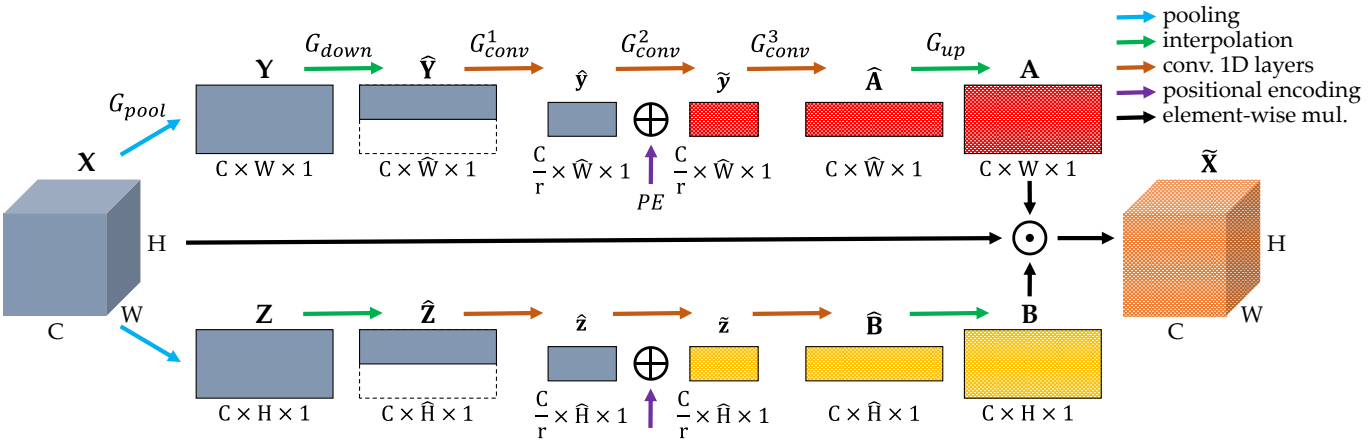


Fig. 2. Architecture of the proposed concurrent horizontal and vertical attention module.

squeeze&excitation for panoramic segmentation. The diagram of the concurrent horizontal and vertical attention module is shown in Fig. 2. To explicitly model the width-wise and height-wise contextual information available in panoramas, the concurrent attention module comprises two branches, each of which adaptively recalibrates the feature maps along channel and the corresponding spatial dimension.

Let $X \in \mathbb{R}^{C\times W\times H}$ denote a feature map in semantic segmentation CNNs, where C is the channel number, W and H are respectively the width and height of the tensor. Given the input feature, the concurrent attention module generates two attention maps $A \in \mathbb{R}^{C\times W}$ and $B \in \mathbb{R}^{C\times H}$. The attention maps are generated through a sequence of operations, which will be introduced in the following paragraphs. The horizontally driven map $A \in \mathbb{R}^{C\times W}$, encodes width-wise contextual information that is distinctly rich in the wide-FoV panoramas. The vertically driven map $B \in \mathbb{R}^{C\times H}$, orthogonally, encodes height-wise structural priors commonly available in street images. After computing the attention maps, the input feature is transformed through the element-wise multiplication of X and the attention maps (A and B), resulting in a recalibrated feature map, formally:

$$\widetilde{X} = X \odot A + X \odot B \tag{1}$$

where $\odot$ denotes element-wise multiplication, $+$ denotes addition of the feature maps. Next, we describe in detail the computation of the width-wise attention map for panoramic street images, while the height-wise attention map, analogously, can be computed along the orthogonal spatial dimension.

Height-Wise Pooling. With the aim of acquiring a channel-wise attention map, in the first step we extract width-wise contextual information from each column by aggregating the input representation X with size $C\times W\times H$ into a matrix Y with the size $C\times W\times 1$ using a height-wise pooling operator G_{pool} (which is implemented by average pooling) to squeeze along the vertical dimension:

$$Y = G_{pool}(X) \tag{2}$$

Precisely, the w-th column vector of Y is calculated as:

$$Y_{:,w} = \left[\frac{1}{H}\sum_{i=1}^{H} X_{1,w,i}; ...; \frac{1}{H}\sum_{i=1}^{H} X_{C,w,i}\right] \tag{3}$$

Interpolation for Coarse Attention. After the height-wise pooling process, a matrix $Y \in \mathbb{R}^{C\times W}$ is produced. However, not all the columns may be necessary for yielding an effective attention map. As illustrated in Fig. 1, class distributions for each direction clearly differ from each other, even if we divide the whole 360° into 18 angles. Thus, we interpolate the $C\times W$ matrix Y into a matrix $\hat{Y}$ with the size $C\times\hat{W}$:

$$\hat{Y} = G_{down}(Y) \tag{4}$$

Correspondingly, in the downstream interpolation, the resulted coarse attention map $\hat{A}$ is transformed to A to have the same width as the input feature map via an upsampling process. Both downsampling and upsampling factors are empirically set equivalent to 4.

Computation of Attention Map. After the squeezing process, we recalibrate the height-wise pooled and interpolated feature map $\hat{Y}$ to provide more importance to relevant horizontal locations and suppress weak ones. Unlike the squeeze&excitation components that use fully connected layers to boost the informative features in classification and pinhole image segmentation tasks [20][21][24], we leverage convolution layers. This is because for panoramic segmentation, we consider that the attention process should allow the relationship between adjacent columns being learned.

Precisely, we adopt three convolution layers to obtain the attention map before upsampling:

$$\hat{A} = \sigma\left[G^3_{conv}\left(\delta\left(G^2_{conv}\left(\delta\left(G^1_{conv}\right)\right)\right)\right)\right] \tag{5}$$

where σ is a sigmoid function and δ is a ReLU activation. The first convolution layer $G^1_{conv}(\hat{Y}) = \hat{y} \in \mathbb{R}^{\frac{C}{r}\times\hat{W}}$ is used for channel reduction as in conventional squeeze&excitation modules which reduces the computation overhead, where r has been set to 4. The second one $G^2_{conv}(\hat{y}) = \tilde{y} \in \mathbb{R}^{\frac{C}{r}\times\hat{W}}$ is applied with sinusoidal positional encoding [27] to the intermediate feature map $\hat{y}$ to better encode the contextual

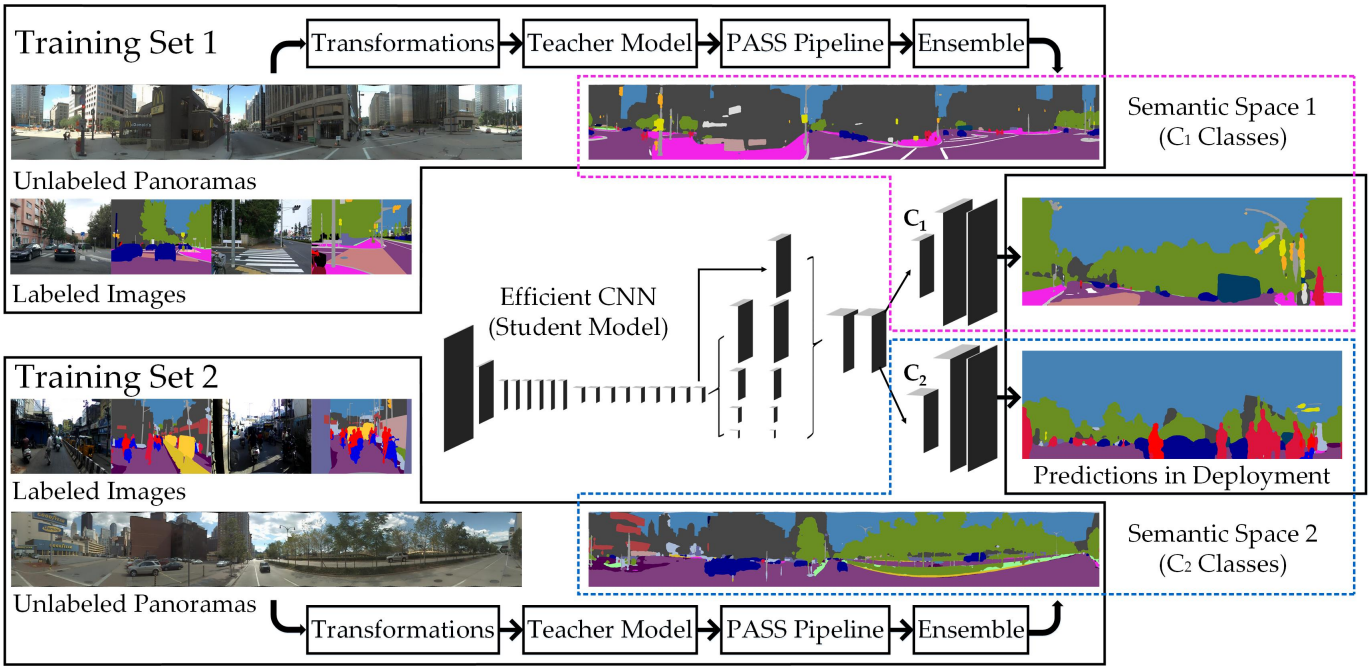


Fig. 3. Overview of the presented multi-source omni-supervised learning scheme for panoramic semantic segmentation. During training, both labeled images and unlabeled panoramas are incorporated. Annotations for unlabeled panoramas are automatically produced by seizing an ensemble of a teacher model's predictions on multiple transformations with the PASS pipeline [9]. During deployment, the student model is not only efficient and suitable for panoramic images, but also robust and capable of outputting multiple sets of visual classes, enriching detectable semantics to fully understand unconstrained surroundings. In the case of two training domains, the efficient CNN is attached with two heads, producing two sets of predictions with C_1 and C_2 classes, respectively.

information through an element-wise addition ($\hat{y} = \hat{y} \bigoplus PE$). Horizontal positions are randomly jittered for improving the generalization capacity to different viewpoints. The last convolution layer $G^3_{conv}(\widetilde{y}) = \hat{A} \in \mathbb{R}^{C \times \hat{W}}$ is adopted to generate the attention map. Finally, we have the width-wise attention map A through the corresponding upsampling:

$$A = G_{up}(\hat{A}) \tag{6}$$

In this way, the recalibration encourages the network to learn more meaningful feature maps that are relevant both channel-wise and spatially (horizontally). The attention map A indicates which channels are critical at each individual column. In other words, for wide-FoV panoramas, it allows to robustly predict correct semantic classes at diverse angles with the correlated channels spotlighted. Similarly, the height-wise attention map B can be computed along the orthogonal (vertical) dimension, which indicates the important channels at each individual row. With their addition as depicted in Equ. (1), our concurrent horizontal and vertical attention module helps to jointly and explicitly exploit width-wise and height-wise contextual cues richly present in panoramic images.

B. Multi-Source Omni-Supervised Learning

For a robust semantic perception of the surroundings in a single pass, it is essential to cover the panoramic domain in the learning stage, i.e., to expose the segmentation network during training to panoramic images which suppose wide FoV and distinct contextual information, both critical to reason holistically about the surroundings. Thereby, we present a multi-source omni-supervised learning scheme for efficient segmentation CNNs, as it is depicted in Fig. 3.

In the preparation phase, we produce annotations for unlabeled panoramas in an automated way by seizing a sophisticated teacher architecture and ensembling the teacher model's predictions, which takes into account the wide-FoV and wrap-around features of panoramic images. In the training stage, the efficient CNN is trained on the union of manually annotated and automatically generated data. In the deployment, the yielded CNN can run in real time due to the student architecture's efficiency, while becoming suitable for panoramic semantic segmentation with rational context awareness.

Preparation Stage. As it is mentioned above, it is important to expose the learner to omnidirectional data in the training. Although large-scale annotated panoramic datasets are not present, there are a battery of panoramas or panoramic videos available in the community. In this work, we leverage a source of unlabeled panoramic images, and automatically create their labels by seizing the known PASS pipeline [9], as shown in Fig. 3. We distill knowledge of semantic segmentation by using a large teacher architecture which may be computationally inefficient that often restricts its application in real-time systems. However, the teacher model's generated semantic maps are high-quality and finely-grained, qualifying its potential usage for data distillation [8]. We ensemble the teacher's predictions on multiple transformed copies of a panorama to produce the final annotation. Specifically, the ensemble procedure takes into account the wide-angle and wrap-around connections of panoramic images.

The teacher model is independently trained on conventional pinhole data. The teacher model F_t, separated into a feature model F_{t_e} that first predicts high-level abstract features and a pixel-wise classification model F_{t_c} that maps the features to

a specific semantic space. When generating the annotations, the panorama I_p (with a size $H_p \times W_p$) is first partitioned into N segments, each of which I_i (size: $H_p \times \frac{W_p}{N}$) is fed into a feature model. This is critical as there is a key correspondence between the features inferred from a panorama segment and the features learned from pinhole images [9], both corresponding to a similar narrow FoV, formally:

$$F_{t_e}\left(\biguplus_{i=1}^{N}(I_i^{H_p\times\frac{W_p}{N}})\right) \equiv F_{t_e}\left(\biguplus_{j=1}^{N_c}(I_{c_j}^{H\times W})\right) \tag{7}$$

where I_{c_j} denotes a conventional image, and $\biguplus$ denotes the concatenation of feature maps.

After the concatenation and a max-pooling operation to recover the feature model size, the classification model F_{t_c} completes the segmentation to yield a pixel-wise semantic map $P_p^{H_p\times W_p}$ for the panorama:

$$P_p^{H_p\times W_p} = F_{t_c}\left[\biguplus_{i=1}^{N} F_{t_e}\left(I_i^{H_p\times\frac{W_p}{N}}\right)\right] \tag{8}$$

This is due to that the classification model with lean convolutional layers, which is also known as the fusion model in the PASS pipeline, is mainly responsible for the classification when the semantically-informative feature map has been already extracted and aggregated. PASS pipeline incurs nearly N times of computation of a single model, as the capacity mostly lies in the feature model. Hence, it is suitable to be used in the preparation rather than the deployment.

Additionally, we seize a specialized ensemble method by taking into account the omnidirectional trait of our task and the wrap-around structures of the panoramas. Concretely, this is realized by rotating the panorama $I_p^{H_p\times W_p}$ for M times along the horizontal direction. In this way, each transformed copy $I_{p_k}^{H_p\times W_p} = \biguplus_{i=1}^{N}(I_{i_k}^{H_p\times\frac{W_p}{N}})$, whose prediction is $P_{p_k}^{H_p\times W_p}$, has a variation of $360°/M$ to the neighboring ones. Then, an ensemble of the predictions can be created, to have the final annotation A_u for an unlabeled panorama, formally:

$$A_u = \bigsqcup_{k=1}^{M} P_{p_k}^{H_p\times W_p} = \bigsqcup_{k=1}^{M} F_{t_c}\left[\biguplus_{i=1}^{N} F_{t_e}\left(I_{i_k}^{H_p\times\frac{W_p}{N}}\right)\right] \tag{9}$$

where $\bigsqcup$ denotes the ensemble process that can be implemented by aggregating the teacher CNN's per-pixel probability maps for the transformed panoramas. An ensemble of predictions is more reliable in nature compared to the single one, since averaging the knowledge of multiple passes makes a model be more prepared against new data. Although this has a direct negative impact in efficiency since having the output of M predictions is inevitably more computation-demanding in running time or memory, these operations (including the PASS pipeline) are processed off-line in a fully automated way and only in the preparation, which will not hurt the efficiency in deployment. Overall, the procedure enables to yield dense and seamless semantic maps more credible for data distillation.

Training Stage. To yield a segmentation model suitable for panoramic images, we propose to seize multiple data sources, as a single training set is limited in the diversity of FoVs, which is prone to overfitting due to all images being gathered with the same camera or certain types of acquisition setup [5]. Precisely, we exploit T large-scale datasets for training, each of which D_i ($i = 1{\sim}T$) corresponds to a specific domain, having labeled samples S_{il}. The annotations for the labeled samples are A_{il}, falling in a semantic class space C_i. To train an efficient student CNN F_s, the conventional strategy is to learn the mapping depicted in the following equation:

$$F_s\left(S_{il}\right) \Longrightarrow A_{il}(C_i) \tag{10}$$

The efficient segmentation model F_s, likewise, can be separated into a student feature model F_{s_e} and a classification model F_{s_c} that maps the predicted features to the specific semantic space, formally:

$$F_s\left(S_{il}\right) = F_{s_c}\left[F_{s_e}\left(S_{il}\right)\right] \Longrightarrow A_{il}(C_i) \tag{11}$$

One of the main aim of our multi-source learning scheme is to train a single CNN (student model) simultaneously in different domains, but the semantic spaces in disparate datasets are often incompatible. For ease of notation, in the case of two domains, $C_1 \neq C_2$, which means that the classes are heterogeneous and class numbers are usually not equivalent, although they are partially overlapping with each other.

However, in spite of the different class definitions, we consider that the relationships encoded in the similar label hierarchies could positively reinforce the generalizability of feature representations when learning across disparate domains. In this sense, it is fruitful to train with multiple datasets, which enables the learner to focus more on essential features. In Fig. 3, the multi-source learning scheme is illustrated in the two domains case, but it can be easily scaled up to multiple domains. To address the heterogeneity in the semantic labels, we append two heads (classification models $F_{s_{c1}}$ and $F_{s_{c2}}$) to the efficient student CNN architecture as it is depicted in Fig. 3, each of which is a fully convolutional module with an upsampling layer for prediction in the specific label space. More precisely, $F_{s_{c1}}$ is responsible for classifying C_1 classes with training set 1, while $F_{s_{c2}}$ is learning for predicting C_2 classes with training set 2. In this way, the training target is modified into:

$$F_{s_{c1}}\left[F_{s_e}\left(S_{1l}\right)\right], F_{s_{c2}}\left[F_{s_e}\left(S_{2l}\right)\right] \Longrightarrow A_{1l}, A_{2l} \tag{12}$$

The domain-specific teacher models, in the preparation stage, have generated two sets of annotations A_{1u} and A_{2u} for the unlabeled panorama samples S_u. Afterwards, the panoramic data in each label space are blended with the pinhole images in that domain for training:

$$F_{s_{ci}}\left[F_{s_e}\left(S_{il}, S_u\right)\right] \Longrightarrow \left(A_{il}, A_{iu}\right)(C_i) \tag{13}$$

In this manner, the student model has been exposed to wide-FoV data in the training. The panoramas will be fed to the learner in both semantic spaces but not necessarily in the same forward/backward passes, which helps to yield more generalized feature representations irrelevant of imagery domains.

Equ. (14) denotes the original prediction bias $Bias_{s_i}$ of the student trained on only pinhole images with model parameter θ_{pin} when it is tested on a panorama sample x_p, whose ground truth is y_{pi} in a specific label space:

$$Bias_{s_i} = \mathbb{E}\left[F_{s_{ci}}\left[F_{s_e}\left(x_p \mid \theta_{pin}\right)\right] - y_{pi}\right] \tag{14}$$

While being omni-supervised, the model parameter $\theta_{pin,pan}$ is learned with both pinhole and panoramic data, where the new prediction bias $Bias'_{s_i}$ can be decoupled into the bias of the prediction to the pseudo label $\hat{y}_{pi}$, and the difference of the pseudo label to the ground truth, as it is shown in Equ. (15). Here, the pseudo label corresponds to the automatically generated labels in the preparation stage.

$$\begin{aligned} Bias'_{s_i} &= \mathbb{E}\left[F_{s_{ci}}\left[F_{s_e}\left(x_p \mid \theta_{pin,pan}\right)\right] - y_{pi}\right] \\ &= \mathbb{E}\left[F_{s_{ci}}\left[F_{s_e}\left(x_p \mid \theta_{pin,pan}\right)\right] - \hat{y}_{pi}\right] \\ &+ \mathbb{E}\left[\hat{y}_{pi} - y_{pi}\right] \\ &\ll Bias_{s_i} + \mathbb{E}\left[\hat{y}_{pi} - y_{pi}\right] \end{aligned} \tag{15}$$

As the student has been exposed to pseudo labels in the training stage, the first term has already been optimized, and thereby it is far smaller than the original bias, as it is depicted in Equ. (15). For the second term, it precisely corresponds to the bias of the teacher with the PASS pipeline [9] which leverages the feature correspondence in a panorama segment semantic segmentation manner as introduced in the preparation stage (see Equ. (7)). Thereby, the second term is also far smaller than the original bias of the student model:

$$\mathbb{E}\left[\hat{y}_{pi} - y_{pi}\right] \equiv Bias_{PASS_{ti}} \ll Bias_{s_i} \tag{16}$$

Thereby, it can be estimated that the prediction bias of the omni-supervised student is far smaller than the model trained with only pinhole images:

$$Bias'_{s_i} \ll Bias_{s_i} \tag{17}$$

In other words, the reliability of the efficient CNN has been significantly improved in panoramic imagery.

Deployment Stage. After training, the student CNN is ready to be taken to open panoramic imagery, while neither ensembling, fusing nor post-processing is required during deployment. The resulted single model, maintaining the efficiency and simplicity as in the common case of a directly supervised end-to-end semantic segmentation approach, also possesses several important benefits. First, as the learner has been exposed to wide-FoV omndirectional and multi-source heterogeneous data, its generalizability has been significantly enhanced in target panoramic domains. Second, it allows the context-aware CNN to leverage the readily accessible contextual information in the full-view 360° image as no separation is incurred in the deployment. Last but not least, the model is able to deliver multiple sets of detectable semantics:

$$\bigvee_{i=1}^{T} F_{s_{ci}}\left[F_{s_e}\left(I_{p_n}\right)\right] = \bigvee_{i=1}^{T}\left(P_i(C_i)\right) \tag{18}$$

where for a new panoramic image I_{p_n}, a union of T predictions of semantic maps will be produced, each of which P_i corresponds to a semantic space C_i, supposing a very rich source of processed information for upper-level navigational applications.

IV. WildPASS Dataset

A. Dataset

To kindle the research on surrounding sensing and to facilitate credible numerical evaluation of semantic segmentation CNNs in panoramic imagery, we put forward the Wild PAnoramic Semantic Segmentation (WildPASS) dataset. As an evaluation dataset, it should be very diverse to reflect the generalizability and real-world applicability of vision algorithms. Unlike mainstream large-scale datasets like Cityscapes [5] and BDD [54] datasets that focus on scene understanding in the urban areas as in Europe or North America, WildPASS embraces the wild by collecting images from all around the world. Precisely, we gather panoramas in 25 cities from all continents (Asia, Europe, Africa, Oceania, North and South America) except for Antarctica.

In Fig. 4, each subfigure gives an annotated panorama example for each city. This is achieved by using the Google Street View and extracting 20 panoramas for each city to form as the full set. We follow the imaging range (360°×70°) of single-shot panoramic annular lens systems installed on instrumented vehicles and robotic platforms [4], by cropping 70° of vertical FoV with the pitch directions from -30° to 40°. We resize all panoramas to 2048×400 pixels.

Table I provides a comprehensive analysis with the most related large-scale databases and evaluation-oriented benchmarks. Our WildPASS possesses several important characteristics among state-of-the-art datasets. It offers a high variability in capturing viewpoint by including both driving scenes and sidewalk environments which has implications for various transportation applications such as automated vehicles and assisted navigation systems for visually impaired people [1]. In addition to viewpoint diversity and global distribution, WildPASS incorporates many unstructured environments with high uncertainties and ambiguities, as well as densely populated areas such as unconstrained traffic intersections that can be imaged in a single 360°. A crucial portion of surroundings in WildPASS are in adverse weather or illumination conditions such as the rainy, snowy and nighttime scenes, which paints a comprehensive picture of real-world challenges for navigational perception systems, expecting a segmentation model that is inherently robust and generalizes well to open, previously unseen domains. Following existing datasets like PASS [9] and RainyNight [61], we create annotations on the most critical navigation-related classes for evaluation. All panoramas from WildPASS are finely labeled by using 6 classes defined in Mapillary Vistas: Car, Road, Sidewalk, Crosswalk, Curb and

Fig. 4. Examples of finely annotated panoramas in WildPASS dataset, collected from 25 cities around the world. From left to right, top to bottom: Adelaide, Beijing, Boston, Budapest, Changsha, Delhi, Hangzhou, Hong Kong, Huddersfield, Johannesburg, Karlsruhe, Kuala Lumpur, London, Los Angeles, Madrid, Manchester, Mumbai, New York, Paris, Reykjavik, Rio de Janeiro, Shanghai, Sydney, Taipei and Tokyo.

TABLE I
ANALYSIS OF RELATED REAL-WORLD DATASETS INCLUDING LARGE-SCALE DATABASES (CITYSCAPES, MAPILLARY VISTAS, BDD10K, IDD20K, WOODSCAPE, TORONTOCITY) AND EVALUATION-ORIENTED BENCHMARKS (WILDDASH, DARKZURICH, RAINYNIGHT, ISSAFE, PASS, WILDPASS).

Dataset	Viewpoint Diversity	Global Distribution	Adverse Scenes	Unstructured Scenes	360°	Number of Images
Cityscapes [5]	✗	✗	✗	✗	✗	5000
Mapillary Vistas [6]	✓	✓	✗	✗	✗	25000
BDD10K [54]	✗	✗	✓	✗	✗	10000
IDD20K [53]	✗	✗	✗	✓	✗	20101
WoodScape [52]	✗	✓	✗	✗	✓	10k
TorontoCity [51]	✗	✗	✗	✗	✓	520
WildDash [60]	✗	✓	✓	✗	✗	211
DarkZurich [50]	✗	✗	✓	✗	✗	151
RainyNight [61]	✗	✗	✓	✗	✗	226
ISSAFE [62]	✗	✗	✓	✓	✗	313
PASS [9]	✓	✗	✗	✓	✓	400
Our WildPASS	✓	✓	✓	✓	✓	500

Person, which are of paramount importance for street scene perception. Compared to previous datasets that have few annotated single-city panoramas [9][51], WildPASS comprises diverse scenes from over 20 cities and multiple continents, encouraging a more realistic assessment of panoramic segmentation performance. Overall, WildPASS has more than 111.5 million annotated pixels for 500 panoramas, which is larger than state-of-the-art evaluation-oriented datasets including WildDash [60], DarkZurich [50], RainyNight [61], ISSAFE [62] and the previous PASS dataset [9].

B. Metrics

To assess whether different context-aware CNNs are ready to sense the surroundings on WildPASS dataset, we use several metrics based on the standard Intersection-over-Union (IoU):

$$\mathrm{IoU} = \frac{TP}{TP + FP + FN} \tag{19}$$

where TP, FP and FN are respectively the number of true positives, false positives and false negatives at the pixel level. The first metric we apply is $\overline{\mathrm{IoU}}$, also known as mean IoU (mIoU), which is the average IoU values for different semantic classes on WildPASS dataset. The second metric is $\widehat{\mathrm{IoU}}$, which is put forward to measure the highest IoU score across different FoVs using cropped panoramas from WildPASS dataset. $\widehat{\mathrm{IoU}}$ denotes the IoU at the most comfortable angle, as it has been demonstrated in [9] that the performance of a semantic segmentation CNN varies for different FoV inputs. The third metric is $\mathrm{p}_{\mathrm{Impact}}$, the impact of 360° FoV on the semantic segmenter, which is calculated as:

$$\mathrm{p}_{\mathrm{Impact}} = \frac{\widehat{\mathrm{IoU}} - \overline{\mathrm{IoU}}}{\widehat{\mathrm{IoU}}} \tag{20}$$

Thereby, a score of 0.0% means that the semantic segmenter perfectly maintains the performance when taken to panoramic imagery, while a value of 50.0% corresponds to a degradation by half of the best accuracy.

V. EXPERIMENTS

A. Datasets

Target Testing Datasets. We perform experiments to evaluate the effectiveness of the proposed concurrent attention module and the omni-supervised training scheme for semantic segmentation in panoramic imagery as in the public PASS [9] (400 images) and the novel WildPASS (500 panoramas) datasets. The PASS dataset was captured by a wearable panoramic annular camera [55] in Hangzhou, China. WildPASS, in contrast, was collected from all around the world. Both of them represent previously unseen, new coming domains, with 6 annotated classes to assess the applicability and generalizability of semantic segmentation CNNs.

Multi-Source Training Sets. Our multi-source training is experimented with two conventional pinhole image datasets: Mapillary Vistas [6] and IDD20K [53], two of the largest street scene parsing datasets in the community. Vistas, provides high diversity with images shot by various cameras across the globe, and the variability of viewpoints with images taken from both

TABLE II
CLASS-WISE ACCURACY OF OMNI-SUPERVISED ERF-PSPNET WITH CONCURRENT HORIZONTAL AND VERTICAL ATTENTION ON MAPILLARY VISTAS. **POL**, **STL**, **BIL** ETC. ARE ABBREVIATIONS OF THE CLASSES. $\overline{\text{IoU}}$: 63.7%.

Pol	**StL**	**Bil**	**TrL**	**Car**	**Tru**	**Bic**	**Mot**	**Bus**	**SiF**	**SiB**	**Roa**	**Sid**
50.1%	30.3%	42.5%	58.4%	90.6%	66.5%	57.1%	55.2%	74.7%	66.6%	31.4%	90.8%	69.9%
Cur	**Fen**	**Wal**	**Bui**	**Per**	**MoC**	**BiC**	**Sky**	**Veg**	**Ter**	**Mar**	**Cro**	$\overline{\text{IoU}}$
58.1%	56.0%	52.4%	86.5%	72.3%	55.6%	52.1%	98.3%	90.0%	67.3%	53.8%	66.1%	63.7%

TABLE III
CLASS-WISE ACCURACY OF OMNI-SUPERVISED ERF-PSPNET WITH CONCURRENT HORIZONTAL AND VERTICAL ATTENTION ON IDD20K DATASET. **ROA**, **DRF**, **SID** ETC. ARE ABBREVIATIONS OF THE CLASSES. $\overline{\text{IoU}}$: 64.6%.

Roa	**DrF**	**Sid**	**NoF**	**Ped**	**Rid**	**Mot**	**Bic**	**AuR**	**Car**	**Tru**	**Bus**	**VeF**
93.3%	62.8%	65.7%	49.2%	66.6%	69.7%	73.9%	41.2%	83.8%	88.1%	81.0%	87.0%	42.7%
Cur	**Wal**	**Fen**	**GuR**	**Bil**	**TrS**	**TrL**	**Pol**	**ObF**	**Bui**	**Bri**	**Veg**	**Sky**
74.5%	55.8%	38.9%	52.0%	60.7%	57.9%	24.2%	48.4%	41.9%	71.7%	64.7%	87.4%	96.7%

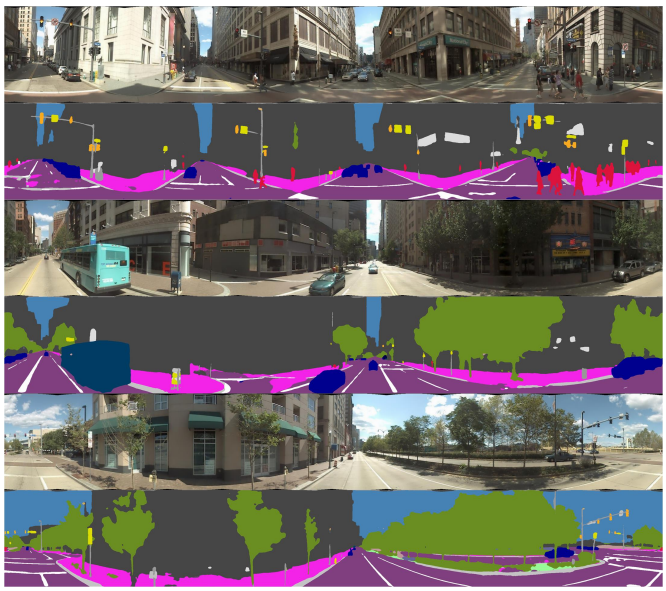


Fig. 5. Examples of automatically generated panoramic annotations for the stitched panoramas using Pittsburgh dataset [63].

perspectives of vehicles and pedestrians. Such variety is critical for panoramic segmentation because it exposes the learner to a wide array of street-scene observations other than only front-facing urban road-driving views. IDD20K, is appealing due to its imported highly unstructured environments that are also implicated in real-world unconstrained surroundings.

Vistas is composed of 18000/2000/5000 images for training/validation/testing. The recently updated IDD20K comprises 14027/2036/4038 images in the train/val/test subsets. Both ground-truth labels in the testing sets are not openly available, but in this research the PASS and WildPASS datasets are readily accessible for evaluation. For Vistas, we use 25 classes for training and present their segmentation accuracy on the validation set in Table II. For IDD20K, we use the level-3 labels (26 classes), where the per-class accuracy in IoU is shown in Table III.

Unlabeled Panoramas Set. As far as the unlabeled panoramas set for data distillation is concerned, we leverage the Pittsburgh dataset [63]. For Pittsburgh dataset, each capture is associated with 24 perspective images with 2 pitch directions and 12 yaw directions. Each perspective image has a horizontal FoV of 60° with overlapping views with the horizontally adjacent ones. By using the known stitching method [64], we stitch the lower pitch images whose perspective matches our target imagery for autonomous navigation. In total, we obtain 966 stitched panoramas from the query set. Fig. 5 shows examples of the stitched panoramas and the generated annotations with our ensemble method. It can be seen that although the labels are not as perfect as manually annotated, they are pretty accurate and well defined. The rich and distinct global contextual information with various directions of roadways and sidewalks being simultaneously imaged, are readily available to be learned by efficient context-aware CNNs through the omni-supervised solution.

B. Training Setups

Teacher Network. We use PSPNet50 [14] as the teacher architecture, which has an $\overline{\text{IoU}}$ of 67.1% on Vistas and 66.5% on IDD20K. Regarding the ensemble, we use the PASS pipeline [9] together with multi-scale prediction, horizontal flipping (mirroring) for PSPNet50 to produce annotations for the unlabeled panoramas. PASS is used with N=4 segments following [9], while the panorama has been rotated for M=32 times which achieves saturated performance. We ensemble the teacher model's predictions by aggregating the probability maps for different copies of a panorama to form as the final label for data distillation. We test the ensemble effects on PASS dataset which shows that PSPNet50 without any operation only achieves 41.4% in $\overline{\text{IoU}}$, while the whole process improves the accuracy to 71.9%. The PASS pipeline plays a vital role as it accounts for an $\overline{\text{IoU}}$ boost of 29.2%. Overall, although the ensemble inevitably increases the time cost with multiple feed-forward passes, it is automatically conducted and only used in the data preparation phase, which does not impair the inference speed of the efficient student CNN to be deployed. With the huge certainty and continuity benefits, the automatically generated panoramic annotations are more credible for data distillation.

Student Network. For the base architecture of the learner CNN, we experiment with ERF-PSPNet [1] due to its real-time performance and publicly available weights of the backbone

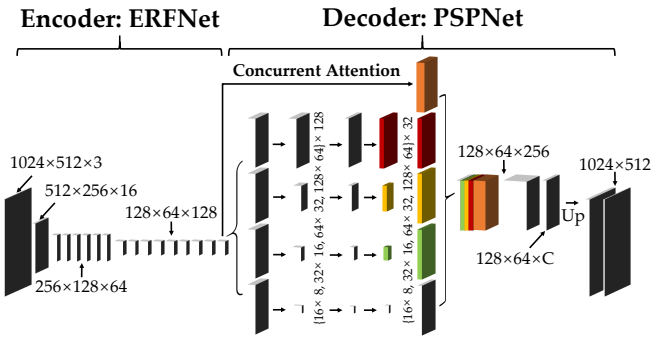


Fig. 6. Architecture of ERF-PSPNet with concurrent horizontal and vertical attention for feature map from encoder and feature maps in pyramid pooling.

pre-trained on ImageNet [65]. As illustrated in Fig. 6, ERF-PSPNet follows an asymmetric encoder-decoder structure, where the encoder is inherited from ERFNet [33] to strike an optimized efficiency-accuracy trade-off, attached by the pyramid pooling module in PSPNet [14] to seize the multi-scale context-sensing capacity. Finally, we use bilinear upsampling to map to the input resolution.

In the omni-supervised solution, our proposed concurrent horizontal and vertical attention module has been inserted in the pyramidal processing before upsampling, which are denoted using different colors for different levels in Fig. 6. The original feature map from the encoder has also been appended with a concurrent attention before being concatenated with the features from the pyramid module. Hence, subsequent convolutions have access to both broad spatial pools and attention driven features. In this way, the network can jointly leverage the rich width-wise and height-wise contextual information available in panoramic images.

Both ERF-PSPNet variants are trained under Adam optimization [66] with a Weight Decay of 2×10^{-4} and an initial Learning Rate of 5.0×10^{-4} that decreases exponentially over 200 epochs. Training samples are fed with a batch size of 12 and a resolution of 1024×512. For the multi-source training, each iteration involves a composition of a forward pass and a backward pass per dataset using cross-entropy loss functions. Class balancing strategy is not used in the loss functions. To focus on studying the effectiveness of multi-source data distillation, we only use random horizontal flipping for data augmentation, where other augmentation techniques [9] that have been proved beneficial for generalization are kept out. In this way, the omni-supervised trained ERF-PSPNet with concurrent attention achieves 63.7% and 64.6% of $\overline{\text{IoU}}$ on Mapillary Vistas and IDD20K, where class-wise accuracies are shown in Table II and Table III. Under the omni-supervised setting for panoramic semantic segmentation, we will also compare our proposal with the independent horizontal attention and the vertical attention, which is separately employed with ERF-PSPNet. Additionally, we will compare with object context estimation and aggregation [28], as well as concurrent spatial and channel 'squeeze & excitation' [21], both similarly employed in each pyramid scale to gather attention maps for context-aware semantic segmentation.

C. Comparison on Public PASS dataset

State of the Art. We step further to evaluate the performance of semantic segmenters in unseen panoramic imagery domains. PASS dataset [9] is a public testbed where many efficient semantic segmentation CNNs have been attempted and experimented by previous researches [4][9]. These efficient CNNs include ENet [31], LinkNet [34], ERF-APSPNet [9], SwaftNet [4], etc., as shown in Table IV. These efficient networks have been trained with an extensive set of data augmentation and style transfer-based domain adaptation strategies [9], which are known beneficial for improving segmentation performance in target imagery.

In this research, we argue that in many real-world applications, no training samples from the target domains will be accessible. Thereby, the evaluation of our proposals follows the paradigm of domain generalization, where one trained model is expected to generalize well in previously unseen scenarios, which is critical for panoramic semantic segmentation in the wild as both the knowledge and image style of the target domains are unforeseeable. In this mode, we also contrast with computation-intensive SegNet [10], PSPNet [14] based on ResNet50 [11], DenseASPP [3] based on DenseNet121 [12] and DANet [26] basd on ResNet50 [11], which are top-accuracy segmentation CNNs, trained on Mapillary Vistas without any domain adaptation by using the proposed hyper-parameters for them in their respective publications.

Baseline. As the annotations of the PASS dataset were created according to the labels definition of Mapillary Vistas, in most cases we evaluate with the Vistas-space result. We could also evaluate by using the IDD20K-space result despite the discrepancy of the classes which may impair the accuracy. As shown in Table IV, when independently-trained, Vistas-based and IDD-supervised ERF-PSPNet achieves 32.2% and 20.1% of $\overline{\text{IoU}}$, respectively. This is much lower than 39.2% which was achieved with aggressive data augmentation which resorts to domain adaptation in the training [9]. Nevertheless, these results show that context-aware CNNs are not ready for sensing the surroundings. This is evidenced not only by our ERF-PSPNet baseline. Even top-accuracy models are rather inaccurate in the panoramic domain due to the large gap between pinhole and 360° imagery. For example, DenseASPP [3] has an $\overline{\text{IoU}}$ of 33.3% while DANet [26] only retains 38.9% when tested against the annular images.

Benefit of Multi-Source Training. Motivated by this issue, our joint-training method, denoted as ERF-PSPNet (IDD20K+Vistas) in Table IV, boosts $\overline{\text{IoU}}$ to 41.0%, significantly higher than independently-yielded scores (20.1% by IDD20K and 32.2% by Mapillary Vistas). The huge gain is attained owing to the well generalized feature representation offered by our multi-source training scheme. Besides, multi-source training offers a diversity of input FoVs with images captured by different cameras, which also helps robustify the segmentation in panoramic imagery. Overall, our multi-source joint-training already achieves a greater $\overline{\text{IoU}}$ (41.0%) than all the best efficient networks attempted by previous researches like ERF-APSPNet (35.5%) and SwaftNet (38.2%).

Benefit of Omni-Supervised Training. Denoted by ERF-

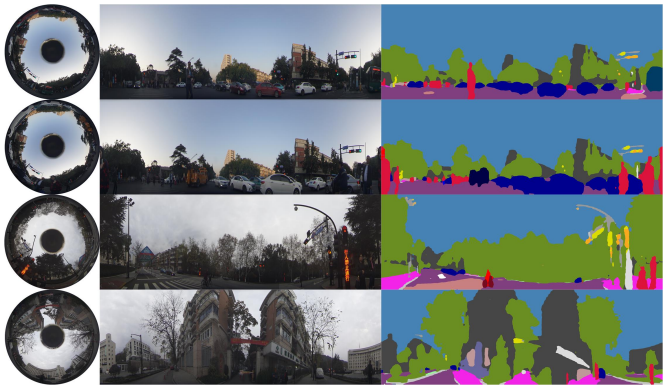
TABLE IV
PER-CLASS ACCURACY ANALYSIS ON PUBLIC PANORAMIC ANNULAR SEMANTIC SEGMENTATION (PASS) DATASET [9]

Network	Car	Road	Sidewalk	Crosswalk	Curb	Person	$\overline{\text{IoU}}$
SegNet [10]	57.5%	52.6%	17.9%	11.3%	11.6%	3.5%	25.7%
PSPNet (ResNet50) [14]	76.2%	67.9%	34.7%	19.7%	27.3%	22.6%	41.4%
DenseASPP (DenseNet121) [3]	65.8%	62.9%	30.5%	8.7%	23.0%	8.7%	33.3%
DANet (ResNet50) [26]	70.0%	67.8%	35.9%	21.3%	12.6%	25.9%	38.9%
ENet [31]	59.4%	59.6%	27.1%	16.3%	15.4%	8.2%	31.0%
LinkNet [34]	62.6%	64.9%	23.2%	6.6%	18.1%	7.5%	30.5%
SQNet [32]	56.5%	57.2%	19.1%	21.4%	10.4%	3.0%	27.9%
ICNet [16]	49.3%	52.4%	20.0%	16.7%	6.7%	9.3%	25.7%
ESPNet [35]	52.6%	51.4%	21.6%	10.5%	6.5%	5.6%	24.7%
EDANet [36]	61.4%	64.0%	28.1%	6.3%	15.2%	8.1%	30.5%
BiSeNet [37]	61.8%	58.3%	17.3%	12.7%	10.8%	5.3%	27.7%
CGNet [38]	65.2%	56.9%	23.7%	3.8%	11.2%	21.4%	30.4%
ERFNet [33]	70.0%	57.3%	25.4%	22.9%	15.8%	15.3%	34.3%
PSPNet (ResNet18) [14]	64.1%	67.7%	31.2%	15.1%	17.5%	12.8%	34.8%
ERF-PSPNet [1]	71.8%	65.7%	32.9%	29.2%	19.7%	15.8%	39.2%
ERF-APSPNet [9]	72.3%	71.4%	32.6%	5.6%	16.3%	14.5%	35.5%
SwiftNet [39]	67.5%	70.0%	30.0%	21.4%	21.9%	13.7%	37.4%
SwaftNet [4]	76.4%	64.1%	33.8%	9.6%	26.9%	18.5%	38.2%
ERF-PSPNet (Vistas)	68.8%	62.0%	26.6%	3.9%	17.5%	14.1%	32.2%
ERF-PSPNet (IDD20K)	53.4%	51.2%	3.2%	0.0%	2.3%	10.6%	20.1%
ERF-PSPNet (IDD20K+Vistas)	**75.5%**	**70.9%**	**32.5%**	**13.0%**	**20.6%**	**33.5%**	**41.0%**
ERF-PSPNet (Omni-Supervised)	**81.4%**	**71.9%**	**39.1%**	**24.6%**	**26.4%**	**44.1%**	**47.9%**
ERF-PSPNet+hA (Omni-Supervised)	82.4%	76.0%	45.9%	20.8%	26.8%	42.6%	49.1%
ERF-PSPNet+vA (Omni-Supervised)	84.0%	74.4%	47.2%	19.8%	30.9%	53.7%	51.7%
ERF-PSPNet+CA (Omni-Supervised)	**85.4%**	**76.5%**	**49.0%**	**27.3%**	**30.3%**	**51.1%**	**53.3%**
ERF-PSPNet+OC [28] (Omni-Supervised)	85.1%	76.8%	41.2%	11.8%	27.9%	53.9%	49.5%
ERF-PSPNet+scSE [21] (Omni-Supervised)	83.3%	75.4%	46.8%	33.3%	28.2%	51.3%	53.0%

PSPNet (Omni-Supervised) in Table IV, the omni-supervised solution further dramatically improves $\overline{\text{IoU}}$ from 41.0% to 47.5%. This even outstrips the large computation-expensive PSPNet (41.4%) by 6.5% and the state-of-the-art DANet (38.9%) by 8.6%. This increase is due to our omni-supervised proposal that exposes the learner to omnidirectional data on the basis of multi-source training, which has been demonstrated to be very important for panoramic semantic segmentation.

Benefit of the Concurrent Attention. Since the omni-supervised solution allows the learner to see panoramas in the training which suppose rich contextual information. We validate the effectiveness of the proposed concurrent attention module in the omni-supervised setting. As shown in Table IV, ERF-PSPNet+CA denotes that with the concurrent attention, it drastically improves $\overline{\text{IoU}}$ to 53.3% from 47.5%. In Table IV, we additionally compare with other attention modules embedded into ERF-PSPNet, including self-attention based object context estimation and aggregation [28] (ERF-PSPNet+OC), as well as concurrent spatial and channel 'squeeze&excitation' [21] (ERF-PSPNet+scSE). For object context processing, it computes a fully dense matrix that measures the similarities of each pixel and each other pixel for the whole feature map, to exploit the indicated degree that they fall in the same semantic class. In this way, ERF-PSPNet+OC models the long-range dependencies. For scSE, it recalibrates the feature maps in spatial and channel dimensions.

However, both OCNet and scSE paths aggregate contextual information without explicit use of the width-wise and height-wise context in panoramic street scenes. Therefore, it turns out ERF-PSPNet+CA outperforms both ERF-PSPNet+OC and ERF-PSPNet+scSE, verifying the superiority of our proposal for panoramic semantic segmentation. Fig. 7 shows segmentation examples produced by our approach on the unfolded annular images from PASS dataset, which demonstrates that 360° seamless segmentation is achievable at the qualified accuracy, without resorting to any panorama separation nor adaptation in the inference that hurts efficiency.



(a) Raw (b) Panorama (c) Prediction

Fig. 7. Qualitative examples of panoramic semantic segmentation on annular images from PASS dataset: (a) Raw, (b) Unfolded panoramas, (c) Predictions.

Ablation of the Concurrent Attention. As shown in Table IV, we conduct an ablation study of the proposed concurrent attention solution (ERF-PSPNet+CA) by also testing with the independent horizontal attention (ERF-PSPNet+hA) and vertical attention (ERF-PSPNet+vA). The horizontal attention encodes width-wise context while the vertical one models height-wise structural priors. It can be seen both attention modules are effective, benefiting the omni-supervised training by explicitly extracting the rich contextual information in panoramas. Finally, the concurrent attention module takes the

TABLE V
PERFORMANCE ANALYSIS ON WILDPASS DATASET.

Network	Accuracy on Vistas	Performance on WildPASS		
	$\overline{\text{IoU}}$	$\overline{\text{IoU}}$	$\widehat{\text{IoU}}$	p_{Impact}
SegNet [10]	51.1%	26.9% (360°)	65.8% (90°)	59.1%
PSPNet (ResNet50) [14]	67.1%	47.3% (360°)	78.7% (100°)	39.9%
DenseASPP (DenseNet121) [3]	65.8%	35.1% (360°)	76.0% (100°)	53.8%
DANet (ResNet50) [26]	62.7%	47.3% (360°)	75.5% (90°)	37.4%
CGNet [38]	53.0%	27.4% (360°)	70.1% (90°)	60.9%
SwiftNet [39]	60.1%	32.9% (360°)	73.4% (90°)	55.2%
SwaftNet [4]	61.8%	36.9% (360°)	75.3% (100°)	51.0%
ERF-PSPNet [1]	61.6%	35.8% (360°)	75.8% (90°)	52.8%
ERF-PSPNet (Vistas+IDD20K)	**63.0%**	**39.4%** (360°)	**75.1%** (110°)	**47.5%**
ERF-PSPNet (Omni-Supervised)	**62.9%**	**58.5%** (360°)	**76.5%** (120°)	**23.5%**
ERF-PSPNet+hA (Omni-Supervised)	63.0%	59.3% (360°)	76.3% (110°)	22.3%
ERF-PSPNet+vA (Omni-Supervised)	63.1%	59.2% (360°)	76.9% (120°)	23.0%
ERF-PSPNet+CA (Omni-Supervised)	**63.7%**	**60.1%** (360°)	**77.4%** (120°)	**22.4%**
ERF-PSPNet+OC [28] (Omni-Supervised)	63.2%	56.9% (360°)	76.7% (110°)	25.8%
ERF-PSPNet+scSE [21] (Omni-Supervised)	63.6%	58.1% (360°)	76.2% (110°)	23.8%

(a) Overall performance analysis and comparison with state-of-the-art accuracy-oriented and efficient networks.

Network	Car	Road	Sidewalk	Crosswalk	Curb	Person	$\overline{\text{IoU}}$
SegNet [10]	58.1%	61.1%	18.2%	4.2%	14.1%	5.9%	26.9%
PSPNet (ResNet50) [14]	81.8%	75.0%	51.8%	23.9%	31.5%	19.9%	47.3%
DenseASPP (DenseNet121) [3]	51.6%	69.2%	38.1%	16.4%	26.3%	8.7%	35.1%
DANet (ResNet50) [26]	81.1%	75.0%	52.9%	30.1%	22.6%	22.1%	47.3%
CGNet [38]	50.2%	60.1%	24.0%	9.9%	15.4%	4.7%	27.4%
SwiftNet [39]	56.5%	64.1%	29.2%	16.2%	22.9%	8.5%	32.9%
SwaftNet [4]	65.2%	68.5%	37.2%	10.7%	26.8%	13.3%	36.9%
ERF-PSPNet [1]	67.9%	70.5%	36.5%	6.4%	24.1%	9.5%	35.8%
ERF-PSPNet (Vistas+IDD20K)	**77.3%**	**73.1%**	**34.5%**	**8.4%**	**21.5%**	**21.9%**	**39.4%**
ERF-PSPNet (Omni-Supervised)	**89.7%**	**79.3%**	**60.9%**	**28.0%**	**38.1%**	**54.8%**	**58.5%**
ERF-PSPNet+hA (Omni-Supervised)	88.2%	81.0%	56.7%	31.8%	40.6%	57.7%	59.3%
ERF-PSPNet+vA (Omni-Supervised)	88.5%	80.2%	59.8%	29.2%	39.9%	57.9%	59.2%
ERF-PSPNet+CA (Omni-Supervised)	**90.2%**	**81.1%**	**59.9%**	**36.0%**	**40.7%**	**52.7%**	**60.1%**
ERF-PSPNet+OC [28] (Omni-Supervised)	90.6%	80.9%	59.0%	16.1%	38.0%	57.0%	56.9%
ERF-PSPNet+scSE [21] (Omni-Supervised)	90.6%	80.1%	59.4%	26.5%	39.7%	52.1%	58.1%

(b) Per-class accuracy analysis on WildPASS dataset.

advantages of both contextual cues in vertical and horizontal directions, reaching an $\overline{\text{IoU}}$ of 53.3% on PASS dataset, which is the new state-of-the-art via a single forward pass.

D. Comparison on WildPASS dataset

Accuracy Downgrade of Context-Aware CNNs. We further conduct comparison on our new WildPASS dataset with the new metrics, which has not been addressed by previous researches. For this reason, we train 4 accurate yet computationally-expensive networks including SegNet [10], PSPNet [14], DenseASPP [3] and DANet [26], as well as 4 efficient networks including CGNet [38], SwiftNet [39], SwaftNet [4] and ERF-PSPNet [1], as shown in Table V. They are trained on Vistas using their optimized hyper-parameters provided by the respective publications. However, when taken to the wild panoramic imagery, the results are disturbing. In Table V, when determining the $\widehat{\text{IoU}}$ at the most comfortable angle of input, we crop different FoVs around the panorama center with variations of 10° for each point and search for the highest IoU value. It can be seen that in spite of the good performance of these context-aware networks in pinhole imagery like the Vistas validation set, and the high accuracy at the most comfortable FoV, they are very sensitive when tested against 360° panoramas from WildPASS and thus suffer from large accuracy downgrades.

To illustrate the impact, DenseASPP degrades by 53.8% as shown in Table V by using the new p_{Impact} metric. For efficient networks like CGNet that stacks a large number of its context guided blocks, suffers from a downgrade of more than 60.0%. This, however, has to be expected, as the use of context, which improves segmentation accuracy, also increases the network's receptive field to attain a global sense of the contextual information. Yet, 360° panoramas feature distinct contextual cues with all horizontal directions being imaged, which is not available in narrow FoVs. Thus, a context-aware CNN with good performance in pinhole data does not necessarily maintain the accuracy and may be more sensitive viewing dissimilar global context in omnidirectional images.

Effectiveness of Multi-Source Omni-Supervision. Motivated by this unsettling sensitivity of segmentation networks, our proposed multi-source omni-supervised solution significantly reduces the impact. As shown in Table V, multi-source training denoted by ERF-PSPNet (Vistas+IDD20K) shifts the most comfortable angle from 90° to 110°. On this basis, the omni-supervised solution denoted by ERF-PSPNet (Omni-Supervised) further shifts the most comfortable FoV to 120° and overall decreases p_{Impact} from 52.8% to 23.5%. The omni-supervised solution slightly decreases the accuracy on Vistas (which is reasonable with only pinhole data), as $\overline{\text{IoU}}$ on WildPASS drastically improves from 39.4% to 58.5%. This demonstrates the effectiveness of multi-source

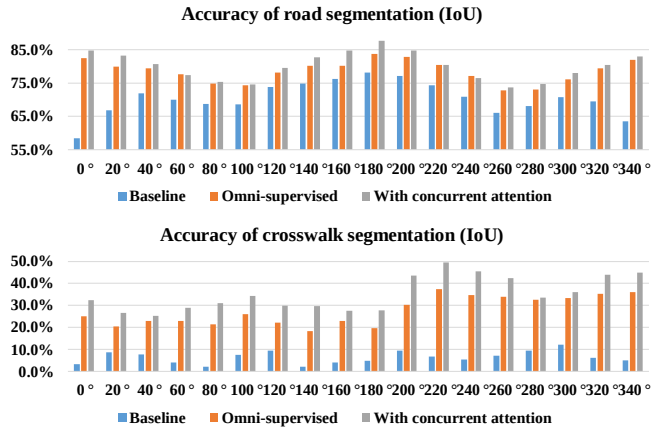


Fig. 8. Accuracy of road and crosswalk segmentation in different directions.

omni-supervision for reducing the gap between pinhole and omnidirectional imagery domains, making efficient CNNs like ERF-PSPNet suitable for ultra wide-FoV inputs.

Fig. 8 further depicts the accuracy at different directions when the 360° is divided into 18 directions. We visualize the performance in IoU of two classes including road and crosswalk, which are the most context-critical classes [9]. Panoramas allow to view multiple directions of roads and crosswalks, which is not available in pinhole images. Thereby, the baseline (vistas-trained model) performs unsatisfactorily where the road segmentation accuracy is much lower at side-view angles (around 90° and 270°) than at forward-view angles (around 180°). Our omni-supervised solution largely improves the performance and yields more uniform accuracy distribution, as the model has learned to segment wide-FoV data with roadways occurring in different directions. For crosswalk, the baseline accuracy is often lower than 10.0%, as generally only one crosswalk region will be observed in pinhole data so crosswalks are frequently labeled as general road markings in panoramic images [9]. In comparison, the omni-supervised solution alleviates this issue and dramatically improves the performance in 360° imagery thanks to our multi-source data distillation.

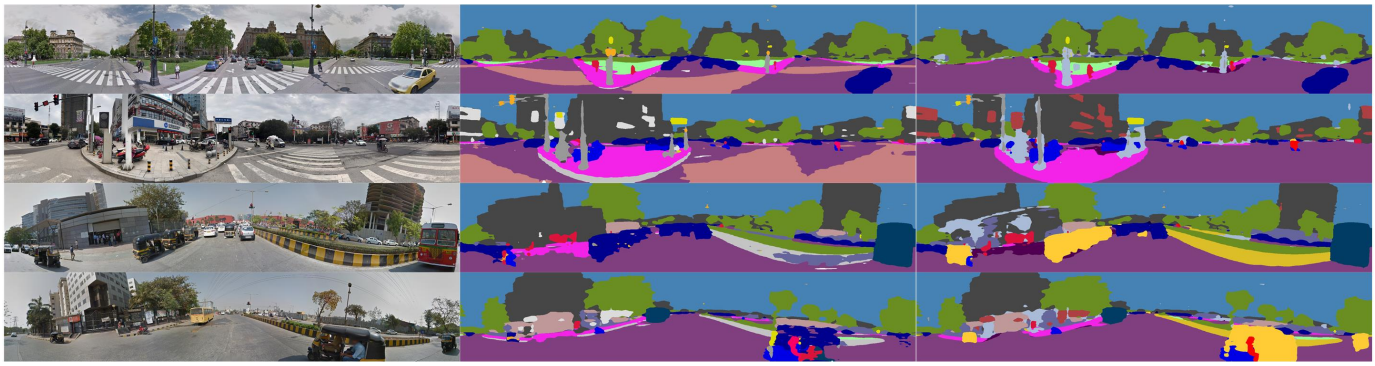
Effectiveness of the Concurrent Attention. When deploying with other attention mechanisms such as OCNet [28] and scSE [21] on WildPASS, it can seen that they decrease the accuracy of omni-supervised ERF-PSPNet. In contrast, our proposed concurrent attention module improves $\overline{\text{IoU}}$ to 60.1%, significantly higher than any other efficient networks, even surpassing those large sophisticated networks like PSPNet and DANet (both 47.3%). In addition, the ablation study on WildPASS verifies that both horizontal and vertical attention are effective, while the concurrent model succeeds to jointly leverage width-wise and height-wise contextual information richly available in panoramic images. Overall, our methods enable the light-weight ERF-PSPNet to clearly stand out in front of state-of-the-art networks, exceeding by 12.8% and 23.2% the best scores achieved by context-aware accuracy-oriented and other efficiency-oriented CNNs (see Table V). As it is shown in Fig. 8, because the concurrent attention allows to learn which channels are critical at each individual column, it consistently improves segmentation accuracy in almost all directions.

TABLE VI
COMPARISON OF OUR CONCURRENT ATTENTION (CA) WITH POSITIONAL ENCODING (PE) AGAINST CONTEXT AGGREGATION METHODS AND OUR MULTI-SOURCE DATA DISTILLATION METHOD AGAINST OTHERS. "OUR CA WITH PE" AND "OUR MULTI-SOURCE + CA" BOTH CORRESPOND TO OUR PROPOSAL: MULTI-SOURCE OMNI-SUPERVISED MODEL WITH CONCURRENT ATTENTION INGRAINED WITH POSITIONAL ENCODING.

Network	Car	Road	Sidew.	Crossw.	Curb	Person	$\overline{\text{IoU}}$
ASPP [13]	88.7%	80.6%	56.4%	18.5%	36.5%	51.4%	55.3%
PPM [14]	89.7%	79.3%	60.9%	28.0%	38.1%	54.8%	58.5%
BFP [19]	89.0%	82.1%	62.1%	18.0%	39.7%	50.3%	56.9%
Strip Pooling [18]	88.7%	81.5%	57.1%	11.3%	39.1%	48.9%	54.4%
Non-local [25]	83.5%	78.4%	52.1%	31.3%	30.0%	46.1%	53.6%
Criss-cross [29]	90.4%	81.6%	61.1%	19.6%	38.5%	54.4%	57.6%
Our CA without PE	90.1%	81.5%	62.3%	26.7%	38.0%	54.6%	58.9%
Our CA with PE	**90.2%**	**81.1%**	**59.9%**	**36.0%**	**40.7%**	**52.7%**	**60.1%**
Standard (Vistas+ADE)	67.8%	71.7%	41.1%	19.8%	22.2%	20.9%	40.6%
Standard (Vistas+IDD)	77.0%	74.4%	33.8%	21.1%	21.5%	17.7%	40.9%
Single-source	91.0%	79.8%	55.1%	10.3%	43.1%	55.0%	55.7%
Our multi-source	**89.7%**	**79.3%**	**60.9%**	**28.0%**	**38.1%**	**54.8%**	**58.5%**
Our multi-source + CA	**90.2%**	**81.1%**	**59.9%**	**36.0%**	**40.7%**	**52.7%**	**60.1%**

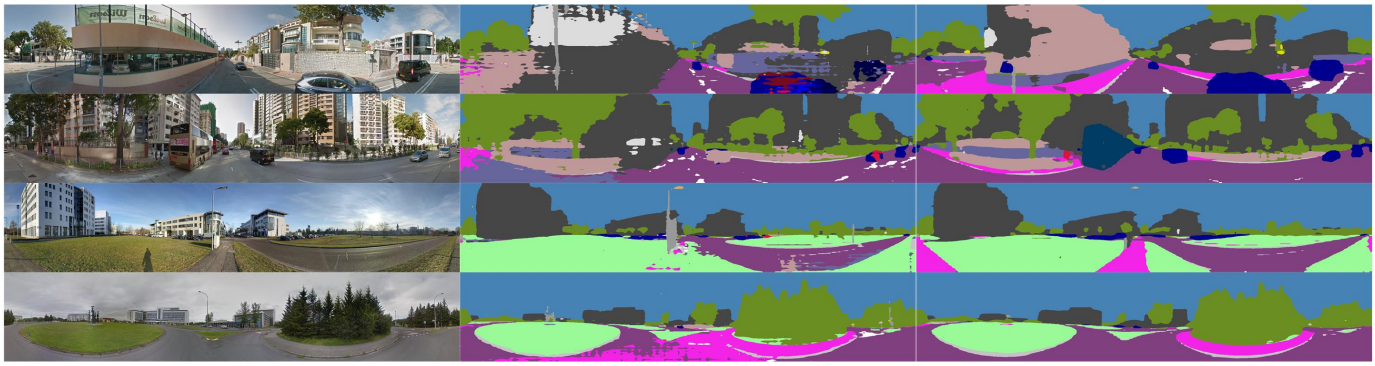
Comparison with Context Aggregation and Distillation Methods. Additionally in Table VI, the proposed concurrent attention is compared against context aggregation methods including Atrous Spatial Pyramid Pooling (ASPP) [13], Pyramid Pooling Module (PPM) [14], Boundary-aware Feature Propagation (BFP) [19], Strip Pooling module [18], non-local module [25] and criss-cross attention [29]. They are both trained with the backbone of ERFNet [33] in a multi-source omni-supervised manner. Due to the learned boundaries, BFP exceeds on the segmentation of roadways and sidewalks, which require accurate road boundary distinction. Criss-cross attention performs well on the detection of cars and persons by leveraging the learned correlations between long-range pixel positions. Overall, while these blocks deliver decent accuracy on some classes, our concurrent attention outperforms them in terms of $\overline{\text{IoU}}$ by clear margins. The results also show that positional encoding improves the generalization, particularly for viewpoint-critical objects like crosswalk.

Furthermore, we compare against standard data-source data distillation methods [8] with ERF-PSPNet by using automatically labeled pinhole images. We consider two settings including the combination of Vistas and ADE20K [67] datasets, and the combination of Vistas and IDD20K. We use unlabeled images from their test sets and generate annotations with the teacher network. Both settings reach similar $\overline{\text{IoU}}$ but much lower than ours. This indicates that the improved performance is not simply due to more training samples but also the knowledge in panoramic images. Especially, our multi-source omni-supervised solution that covers omnidirectional data in the training, helps materialize the knowledge for context-ware panoramic segmentation. Finally, when our solution is operating in the single-source omni-supervised manner, i.e., only with Vistas-space data, it can reach high segmentation accuracy for common object classes like cars and persons. In contrast, our multi-source solution excels at identifying challenging stuff classes including sidewalks and crosswalks which exhibit higher diversity in real-world surroundings,



(a) Panorama (b) Prediction in Vistas space (c) Prediction in IDD20K space

Fig. 9. Qualitative examples of panoramic semantic segmentation on panoramas from WildPASS dataset: (a) Panoramas, (b) Predictions in Vistas space and (c) Predictions in IDD20K space.



(a) Panorama (b) DenseASPP [3] (c) Ours

Fig. 10. Qualitative examples of panoramic semantic segmentation on panoramas from WildPASS dataset predicted by our approach compared with the state-of-the-art: (a) Panoramas, (b) Predictions of DenseASPP [3] and (c) Our omni-supervised ERF-PSPNet with concurrent attention.

while only slightly falling behind for those common objects. Thereby, our multi-source solution surpasses the single-source omni-supervised method by a margin of 2.8% in $\overline{\text{IoU}}$, which indicates that the increased diversity is essential for robust segmentation in the wild.

E. Qualitative Analysis and Discussion

Prediction in Multiple Spaces. Fig. 9 displays representative predictions of our omni-supervised ERF-PSPNet with concurrent attention in multiple semantic spaces on panoramas from WildPASS dataset. On the one hand, clear and highly robust segmentation is achieved in the unseen panoramic domains. Besides, in this demonstration, it is shown that while only a single model is yielded, it delivers multiple sets of visual classes that are complementary to each other. For instance, as it can be seen from Fig. 9b, crosswalks and curbs are predicted in the Vistas space which are absent in IDD20K. In comparison, IDD20K-space results can help to foresee safety-critical classes like auto-rickshaws ("tuk-tuk" vehicles) whose behavior is highly unpredictable, as shown in Fig. 9c (denoted in yellow). As a result, the detectable semantics and recognizable classes have been enriched, which are often required and useful for upper-level applications to gain a complete scene comprehension, especially in real-world unconstrained surroundings.

TABLE VII
SPEED ANALYSIS IN FRAMES PER SECOND (FPS) FOR DENSEASPP, DS-PASS AND ERF-PSPNET VARIANTS ON VARIOUS GPU PROCESSORS.

Network	Speed on GPU processors		
	Titan RTX	1080Ti	2080Ti
DenseASPP (DenseNet121) [3]	70.6	30.8	57.7
DS-PASS (SwaftNet) [4]	209.9	76.3	56.8
ERF-PSPNet	164.7	91.2	132.1
ERF-PSPNet+OC [28]	121.3	72.0	87.6
ERF-PSPNet+scSE [21]	139.5	74.3	115.0
ERF-PSPNet+CA	122.2	77.0	97.0

Qualitative Comparison. Fig. 10 shows the segmentation results of our solution in comparison with one of the state-of-the-art accurate network DenseASPP [3]. It can be clearly seen that DenseASPP with a large receptive field is vulnerable when taken to the panoramic imagery in the wild. Consequently, it produces unstable semantic maps due to inappropriate context awareness. In contrast, our solution enables the efficient ERF-PSPNet to reliably leverage the contextual information in panoramas thanks to both the omni-supervision and concurrent attention. Thereby, it yields high-quality semantics and consistent roadways segmentation, even if they feature distinct global distribution to those in pinhole images.

Speed Analysis. In Table VII, we report the speed measured in Frames Per Second (FPS) of ERF-PSPNet variants at the input resolution of 1024×512. The FPS metric directly

corresponds to the processing time tested on different GPU processors including NVIDIA Titan RTX, GTX 1080Ti and RTX 2080Ti, where the batch size has been set to 1 to simulate real-time applications. Following [9], we report the mean FPS results over 400 forward passes running through all panoramas in the PASS dataset. In Table VII, we compare with DenseASPP [3], a state-of-the-art accurate network while requiring merely moderate computation power, and DS-PASS [4], a panoramic segmentation approach with one of the best efficient architecture SwaftNet. While DS-PASS is very fast on Titan RTX with sufficient computation budget, it slows down on other cost-effective GPUs, which is due to the separation and fusion process which significantly burdens the deployment. In contrast, from Table VII, it can be seen that the high inference speed of ERF-PSPNet is maintained at the level far above the real-time constraint. As another illustration of this superiority, PASS only runs at 40.2FPS (reported in [9]) with the same ERF-PSPNet architecture on NVIDIA Titan RTX, which is clearly slower than our approach (more than 120.0FPS). This is because with our omni-supervised solution, no panorama separation nor domain adaptation is incurred in the deployment, while the proposed concurrent attention only marginally increases the computational complexity.

VI. Conclusion

In this paper, we have addressed panoramic semantic segmentation in the wild. We have observed that state-of-the-art CNNs such as PSPNet and DANet, despite their wide-range context sensing capacity which improves their accuracy in pinhole data, suffer from large performance deterioration when taken to panoramic imagery which features distinct global contextual information. To explicitly exploit the width-wise and height-wise contextual cues markedly available in wide-FoV panoramas, we have presented a concurrent horizontal and vertical attention module. The potential of rich global context information is unlocked also thanks to our designed multi-source omni-supervised learning scheme, which covers the panoramic imagery in the training of efficient CNNs.

We have put forward the WildPASS dataset, with panoramas collected from all around the world embracing challenging unconstrained surroundings, to facilitate credible evaluation of semantic segmentation CNNs in panoramic imagery. An extensive set of experiments demonstrates the effectiveness of our proposals, which enable high-efficiency architectures like ERF-PSPNet to attain significant generalization benefits in open panoramic domains, outperforming the state-of-the-art on both the public PASS and the novel WildPASS datasets.

In the future, we will expand the WildPASS dataset by collecting world-wide panoramas from more cities and creating annotations on more classes like structured objects including truck and bus. We have the intention to explore uncertainty-inspired multi-space fusion and efficiency-oriented non-local modules to capture omni-range contextual dependencies. We aim to fulfill panoramic panoptic segmentation and change detection to enable a more unified and comprehensive scene understanding for autonomous navigation 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)*, 2018, pp. 1033–1038.

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

[3] 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*, 2018, pp. 3684–3692.

[4] K. Yang, X. Hu, H. Chen, K. Xiang, K. Wang, and R. Stiefelhagen, "Ds-pass: Detail-sensitive panoramic annular semantic segmentation through swaftnet for surrounding sensing," in *2020 IEEE Intelligent Vehicles Symposium (IV)*, 2020, pp. 186–193.

[5] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3213–3223.

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

[7] S. Choi, J. T. Kim, and J. Choo, "Cars can't fly up in the sky: Improving urban-scene segmentation via height-driven attention networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9373–9383.

[8] I. Radosavovic, P. Dollár, R. Girshick, G. Gkioxari, and K. He, "Data distillation: Towards omni-supervised learning," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4119–4128.

[9] K. Yang, X. Hu, L. M. Bergasa, E. Romera, and K. Wang, "Pass: Panoramic annular semantic segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 10, pp. 4171–4185, 2020.

[10] 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.

[11] 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)*, 2016, pp. 770–778.

[12] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269.

[13] 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.

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

[15] H. Ding, X. Jiang, B. Shuai, A. Q. Liu, and G. Wang, "Context contrasted feature and gated multi-scale aggregation for scene segmentation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2393–2402.

[16] 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.

[17] H. Ding, X. Jiang, B. Shuai, A. Q. Liu, and G. Wang, "Semantic segmentation with context encoding and multi-path decoding," *IEEE Transactions on Image Processing*, vol. 29, pp. 3520–3533, 2020.

[18] Q. Hou, L. Zhang, M.-M. Cheng, and J. Feng, "Strip pooling: Rethinking spatial pooling for scene parsing," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4003–4012.

[19] H. Ding, X. Jiang, A. Q. Liu, N. M. Thalmann, and G. Wang, "Boundary-aware feature propagation for scene segmentation," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6818–6828.

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

[21] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel 'squeeze & excitation' in fully convolutional networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2018, pp. 421–429.

[22] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 510–519.

[23] H. Zhang *et al.*, "Context encoding for semantic segmentation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7151–7160.

[24] X. Hu, K. Yang, L. Fei, and K. Wang, "Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 1440–1444.

[25] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.

[26] J. Fu *et al.*, "Dual attention network for scene segmentation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3141–3149.

[27] A. Vaswani *et al.*, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

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

[29] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 603–612.

[30] S. Huang *et al.*, "Ordnet: Capturing omni-range dependencies for scene parsing," *IEEE Transactions on Image Processing*, vol. 29, pp. 8251–8263, 2020.

[31] 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.

[32] M. Treml *et al.*, "Speeding up semantic segmentation for autonomous driving," in *MLITS, NIPS Workshop*, vol. 2, 2016, p. 7.

[33] 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.

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

[35] 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.

[36] S.-Y. Lo, H.-M. Hang, S.-W. Chan, and J.-J. Lin, "Efficient dense modules of asymmetric convolution for real-time semantic segmentation," in *Proceedings of the ACM Multimedia Asia on ZZZ*, 2019, pp. 1–6.

[37] 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.

[38] T. Wu, S. Tang, R. Zhang, J. Cao, and Y. Zhang, "Cgnet: A light-weight context guided network for semantic segmentation," *IEEE Transactions on Image Processing*, vol. 30, pp. 1169–1179, 2020.

[39] M. Oršic, I. Krešo, P. Bevandic, and S. Šegvic, "In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 12 599–12 608.

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

[41] 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.

[42] 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 *2017 IEEE International Conference on Intelligent Transportation Systems (ITSC)*, 2017, pp. 1–8.

[43] 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)*, 2018, pp. 132–137.

[44] 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)*, 2018, pp. 1767–1772.

[45] C. Zhang, S. Liwicki, W. Smith, and R. Cipolla, "Orientation-aware semantic segmentation on icosahedron spheres," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 3532–3540.

[46] I. Budvytis, M. Teichmann, T. Vojir, and R. Cipolla, "Large scale joint semantic re-localisation and scene understanding via globally unique instance coordinate regression," in *British Machine Vision Conference (BMVC)*, 2019, p. 31.

[47] Y. Xu, K. Wang, K. Yang, D. Sun, and J. Fu, "Semantic segmentation of panoramic images using a synthetic dataset," in *SPIE*, vol. 11169, 2019, p. 111690B.

[48] A. R. Sekkat, Y. Dupuis, P. Vasseur, and P. Honeine, "The omniscape dataset," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 1603–1608.

[49] 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)*, 2019, pp. 1312–1318.

[50] C. Sakaridis, D. Dai, and L. Van Gool, "Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 7373–7382.

[51] S. Wang *et al.*, "Torontocity: Seeing the world with a million eyes," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3028–3036.

[52] S. Yogamani *et al.*, "Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9307–9317.

[53] 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)*, 2019, pp. 1743–1751.

[54] F. Yu *et al.*, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2636–2645.

[55] K. Yang *et al.*, "Can we pass beyond the field of view? panoramic annular semantic segmentation for real-world surrounding perception," in *2019 IEEE Intelligent Vehicles Symposium (IV)*, 2019, pp. 446–453.

[56] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[57] Q. Guo *et al.*, "Online knowledge distillation via collaborative learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 020–11 029.

[58] L.-C. Chen *et al.*, "Naive-student: Leveraging semi-supervised learning in video sequences for urban scene segmentation," in *European Conference on Computer Vision*, 2020, pp. 695–714.

[59] Y. Zhu *et al.*, "Improving semantic segmentation via self-training," *arXiv preprint arXiv:2004.14960*, 2020.

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

[61] S. Di *et al.*, "Rainy night scene understanding with near scene semantic adaptation," *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[62] J. Zhang, K. Yang, and R. Stiefelhagen, "Issafe: Improving semantic segmentation in accidents by fusing event-based data," *arXiv preprint arXiv:2008.08974*, 2020.

[63] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi, "Visual place recognition with repetitive structures," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 883–890.

[64] M. Zhu, W. Wang, B. Liu, and J. Huang, "A fast image stitching algorithm via multiple-constraint corner matching," *Mathematical Problems in Engineering*, vol. 2013, pp. 1–6, 2013.

[65] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[66] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015, pp. 1–15.

[67] 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)*, 2017, pp. 5122–5130.