Omnisupervised Omnidirectional Semantic Segmentation

Kailun Yang[1], Xinxin Hu[2], Yicheng Fang[2], Kaiwei Wang[2], and Rainer Stiefelhagen[1]

***Abstract*—Modern efficient Convolutional Neural Networks (CNNs) are able to perform semantic segmentation both swiftly and accurately, which covers typically separate detection tasks desired by Intelligent Vehicles (IV) in a unified way. Most of the current semantic perception frameworks are designed to work with pinhole cameras and benchmarked against public datasets with narrow Field-of-View (FoV) images. However, there is a large accuracy downgrade when a pinhole-yielded CNN is taken to omnidirectional imagery, causing it unreliable for surrounding perception. In this paper, we propose an omnisupervised learning framework for efficient CNNs, which bridges multiple heterogeneous data sources that are already available in the community, bypassing the labor-intensive process to have manually annotated panoramas, while improving their reliability in unseen omnidirectional domains. Being omnisupervised, the efficient CNN exploits both labeled pinhole images and unlabeled panoramas. The framework is based on our specialized ensemble method that considers the wide-angle and wrap-around features of omnidirectional images, to automatically generate panoramic labels for data distillation. A comprehensive variety of experiments demonstrates that the proposed solution helps to attain significant generalizability gains in panoramic imagery domains. Our approach outperforms state-of-the-art efficient segmenters on highly unconstrained IDD20K and PASS datasets.**

***Index Terms*—Intelligent Vehicles, Scene Understanding, Semantic Segmentation, Scene Parsing, Omnisupervised Learning, Omnidirectional Images.**

I. Introduction

THE breakthrough of Convolutional Neural Networks (CNNs) has greatly advanced the frontiers of computer vision algorithms, as in image classification [1], cropping [2], segmentation [3] and tracking [4][5]. Vision-based semantic segmentation unifies typically separate detection tasks by rendering a pixel-wise scene understanding [6]. It allows to solve many problems at once and therefore has been allied to the perception in Intelligent Vehicles (IV) [7]. CNNs excel at this task due to the development of deep architectures [8][9] and the emergence of large datasets [10][11]. Modern efficient networks become capable of performing road-driving scene semantic segmentation both swiftly and accurately [12][13][14],

Manuscript received February 26, 2020; revised May 21, 2020; revised July 23, 2020; accepted September 8, 2020. This work was supported in part by Federal Ministry of Labor and Social Affairs (BMAS) through the AccessibleMaps project under the grant number 01KM151112, in part by Hangzhou SurImage Technology Company Ltd., in part by Hangzhou Huan-Jun Technology Company Ltd., and in part by Hangzhou KrVision Technology Company Ltd. (krvision.cn). *(Corresponding author: Kailun Yang.)*

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

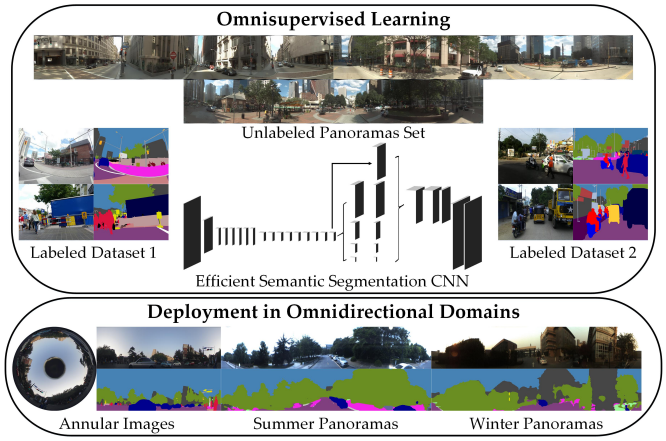


Fig. 1. Overview of the proposed omnisupervised solution: an efficient CNN is trained using multi-source labeled images and unlabeled panoramas, yielding a single model suitable for omnidirectional road sensing domains.

which provides a rich resource of processed high-level information for upstream navigational applications.

At the same time, omnidirectional images are omnipresent in IV systems thanks to their complete surrounding sensing capacity. However, most of the current semantic segmenters are predominantly designed to work with forward-facing cameras [10][15], which heavily limits the Field of View (FoV) and the sufficiency of acquired information. One of the essential reasons lies in that most publicly available datasets merely contain pinhole images. Nevertheless, when taking a pinhole-yielded CNN to omnidirectional imagery, the performance drops significantly and even catastrophically [14], causing it unreliable for the perception of the whole surroundings.

While the straightforward solution is to create a large-scale surround-view dataset for fully supervised training, the ground-truth acquisition entails extremely time-consuming and error-prone labeling procedures. It is particularly exacerbated for omnidirectional data, where human annotation is highly prohibitive [16]. There are a few wide-angle datasets [17][18][19], but their scene diversity and annotation density are far lower than popular pinhole databases [10][11]. This is due to the novelty of omnidirectional sensors as well as the higher complexities and distortions implicated in wide-FoV images [20][21]. It motivates a subset of research works to synthesize omnidirectional data from conventional pinhole segmentation sets [22][23][24]. Nevertheless, the large discrepancy between synthetic and real-world imagery sparks further domain adaptation methods [25][26][27], which require having access to images from a specific target environment,

but struggle to generalize in open panoramic domains. Another cluster of methods [7][14][16], with the aim of reusing knowledge learned from pinhole data, separates the panorama into several segments for semantic maps prediction and fusion. However, this induces significant computation complexity [16]. Besides, while panoramas allow to analyze the whole scene with all the context available, these solutions fail to leverage the global contextual cues with the panorama partitioned into discrete segments.

To address these issues, we propose an omnisupervised learning framework for efficient semantic segmentation CNNs, which bridges multiple heterogeneous data sources (see Fig. 1). Being omnisupervised, the efficient CNN exploits both labeled pinhole images and unlabeled 360° full-view panoramas. Precisely, the omnisupervised learning concept is approached through data distillation [28], where we distill knowledge of a teacher model by creating an ensemble of its predictions run on multiple transformations of the panoramas. We put forward a specialized ensemble method for panoramas by taking into consideration their wrap-around connections along with the Panoramic Annular Semantic Segmentation (PASS) pipeline [16], which facilitates the generation of well-defined omnidirectional labels, bypassing the laborious per-pixel manual annotation process.

The proposed omnisupervised solution produces a single student model trained on the union of manually annotated and automatically generated data. The designed multi-source training helps the learner to attain significant generalizability benefits when taken to unseen domains. Moreover, while yielded as a single segmenter, it is able to deliver outputs in multiple semantic spaces, enriching the recognizable classes required to fully understand real-world unconstrained surroundings. As the efficient CNN has been exposed to omnidirectional data in the training stage, it is directly suitable in panoramic imagery without any adaptation/separation, while retaining access to the crucial global contextual information.

A comprehensive variety of experiments is conducted with our high-efficiency ERF-PSPNet [6], while we study in a general way that is applicable to any efficient architecture. Our approach outperforms previous state-of-the-art efficient segmenters on the challenging IDD20K [15] and PASS [16] datasets. The omnisupervised CNN is further deployed on an instrumented vehicle, where we collect panoramic images under various weather and illumination conditions, as well as in different cities, verifying the generalization capacity of our solution. Our datasets and codes are open-sourced at.[1]

II. Related Work

A. Efficient Semantic Segmentation

Fully Convolutional Networks (FCNs) [8] started the era of end-to-end semantic segmentation, whose performance was outsripped by SegNet [29], DRNet [30], PSPNet [31] and DeepLab [32]. A host of modules have been adopted to aggregate the crucial contextual information. PSPNet applies a Pyramid Pooling Module (PPM) to capture multi-scale context, which is similarly achieved by using Atrous Spatial Pyramid Pooling (ASPP) as introduced in DeepLab. DenseASPP [33] extends ASPP by incorporating dense connections. DANet [34] captures long-range dependencies via a self-attention mechanism while OCNet [35] harvests object context. ACNet [36] exploits complementary features in an attention-bridged way. In [37], the semantic granularity gap is alleviated to improve the fusion of shallow and deep features. These works have achieved high-quality segmentation on existing pinhole benchmarks. However, the developed networks are computationally-intensive, limiting the deployability in response time-critical IV applications.

To achieve real-time inference, enormous efforts have been paid by proposing efficient networks like ENet [38], ERFNet [12], LinkNet [39], SQNet [40], ICNet [41], ESP-Net [42], EDANet [43], BiSeNet [44], CGNet [45], ERF-PSPNet [6], ERF-APSPNet [16], SwiftNet [13] and Swaft-Net [14]. Specifically, ENet and ERFNet follow asymmetric encoder-decoder structures with early downsampling, while ICNet and BiSeNet are built on multi-branch setups. In LinkNet and SwiftNet, the compact U-shape architectures are completed with skip connections and ladder-style upsampling, while CGNet is packed with context guided blocks in all stages. SwaftNet [14] is crafted for high-resolution data like panoramas with squeeze-and-excite [46] in the lateral connections to enhance detail sensitivity. Nevertheless, most efficiency-oriented networks are designed to solve the trade-off between efficiency and accuracy in conventional imagery, while the yielded segmenters suffer from significant accuracy drops when taken to omnidirectional domains [16].

B. Omnidirectional Semantic Segmentation

To enlarge the FoV of semantic perception, researchers proposed to use fisheye sensors [21][24] or install an array of cameras as a surround-view platform [20][47]. This often entails the use of multiple cameras but the number of devices is one of the most critical parameters to be optimized for IV systems. Zhang et al. [22] proposed to use a single spherical input and converted it to an unfolded icosahedron mesh for a holistic labeling of the surroundings. They created an Omni-SYNTHETIA dataset which was produced from the virtual road-driving SYNTHIA dataset [48]. Sharing a similar spirit, Xu et al. [23] introduced a SYNTHIA-PANO dataset by stitching different directions of synthetic observation into a panoramic image. In this line, Budvytis et al. [19] performed joint semantic scene understanding and localization with a CamVid-360 panoramic dataset, which was collected by cycling along the original path of the well-known CamVid database [49]. The models in these works were trained with their omnidirectional data, which are far less realistic nor diverse than large-scale pinhole natural image sets [11].

In contrast, Yang et al. [7][16] proposed a Panoramic Annular Semantic Segmentation (PASS) framework. Unlike approaches that are dependent on omnidirectional labeled data, they trained on conventional images and deployed in unseen

[1]Datasets and Codes for Omnisupervised Omnidirectional Semantic Segmentation: https://github.com/elnino9ykl/OOSS

panoramas. They used a distortion-controlled panoramic annular camera for their navigation assistance application, where 360° semantic segmentation was predicted in a single process on the unfolded panorama. This paradigm unlocked the use of panoramic sensing system in a great variety of scenarios by taking the advantage of the wealth of pinhole image datasets. They further applied PASS to support visual odometry and proposed SwaftNet [14] for detail-aware driving scene parsing. In spite of being deployed with efficient CNNs, running time is significantly higher than that of end-to-end prediction [16], as the panoramic maps are first separated and then fused in their PASS pipeline. As another consequence, the segmenter fails to exploit the crucial global contextual information with the single-shot panorama partitioned into pieces for several forward passes. Unlike previous works, we aim to produce a single model directly applicable in omnidirectional imagery without any separation/adaptation that hurts inference efficiency. Our omnisupervised learning framework, which covers panoramic imagery in training, is orthogonal to the prior works that focus on the adaptation in deployment.

C. Knowledge Distillation and Domain Adaptation

CNN architectures have a high reliance on large-scale visual data. To address the lack of annotated data and boost the efficiency in vision systems, knowledge distillation has been introduced to transfer information from one model to another [50], e.g., from an accurate yet computation-intensive model to a fast one. Thereby, knowledge distillation is often characterized as a teacher-student training framework. It has been widely investigated in classification [51][52][53], tracking [54] and re-identification [55] tasks. In [50], a systematic analysis of knowledge distillation methods was provided. Here, we mainly review some related knowledge distillation methods on the unification of multi-source supervision.

In [52], distillation was formalized to unify heterogeneous classifiers from multiple sources that cannot directly share their data, which allows to transfer the knowledge without requiring the target classes of all teacher and student models to be the same. Such class contradiction problems exist in semantic segmentation, which are usually addressed by regulating the complex class hierarchies among datasets [56]. In [53], a self-paced distillation scheme was designed to aggregate knowledge from multiple experts that are learned on less imbalanced subsets of the entire long-tailed distribution. While it facilitates to yield a unified student model, retraining of multiple large teacher networks is necessitated. In [54], mutual learning was enabled through a multi-students learning mechanism with knowledge shared between students. In [57], a relational knowledge distillation method was presented to transfer mutual relations of data examples. Recently, knowledge distillation has also been extended to semantic segmentation with the aim of training compact networks with the help of cumbersome networks [58][59][60]. For example, local pixel-wise probabilities were mainly distilled in [58] to transfer the knowledge gained by a heavy network to guide the learning of fast networks.

Our omnisupervised solution also pursues multi-source distillation but we aim to improve the reliability of a single student network across domains, i.e., from pinhole to panoramic imagery. Unlike previous model distillation methods that ensemble multiple experts but entail re-training different heavy networks, we perform data distillation [28] with a single teacher architecture, arguably more flexible in the large database case. We avoid the complex regulation of the class conflictions between datasets by appending multiple heads and using the unlabeled panoramas as a bridge. Compared to the data distillation work for human keypoint detection that requires huge amounts of extra generated data [28], we only use a moderate amount of unlabeled panoramas, while largely robustifying omnidirectional semantic segmentation. In our omnisupervised learning system, the knowledge is distilled from a light-weight ensemble formed by multiple data transformations considering the wide-angle and wrap-around connections of panoramic images. In addition, our framework, being simple and effective, does not require to drastically modify network structures or impose consistency by adding any extra loss terms.

There is also a great volume of works on domain adaptation [25][26][27][61][62][63] for semantic segmentation. In [25][27], image-level style transfers were performed to bridge the day and night domain gap. In [26], a battery of input adapters was used depending on the weather conditions. These largely sacrifice the inference efficiency of semantic perception frameworks when deployed in IV applications. Another trend of methods [61][62][63] reduced the gap between the feature distributions across domains without the need of generating new images in the testing. However, they require having access to a large set of images from the specific target domain during training. Contrarily, our omnisupervised solution operates in the domain generalization paradigm by using a set of unlabeled panoramas irrelevant of the target domains. We expect the yielded model to generalize to panoramic imagery of open domains and previously unseen environments.

III. Framework

The diagram of the proposed omnisupervised learning framework for omnidirectional semantic segmentation is depicted in Fig. 2. In the preparation stage, we generate annotations for unlabeled panoramas by using a large teacher architecture and ensembling the teacher model's predictions. In the training phase, we blend multi-source pinhole images with manually annotated labels and panoramas with automatically generated labels to train the student network. In the deployment phase, the yielded CNN can run in real time due to the student model's efficiency, while becoming suitable for omnidirectional semantic segmentation. In the following subsections, we describe in detail the preparation (Sec. III-A), training (Sec. III-B) and deployment (Sec. III-C) of efficient CNNs for omnisupervised omnidirectional semantic segmentation. In Sec. III-D, we describe the efficient learner ERF-PSPNet, which is used as the student model illustrated in Fig. 2 and its architecture variants.

A. Preparation Stage

To achieve robust omnidirectional semantic segmentation, it is important to expose the model to omnidirectional data

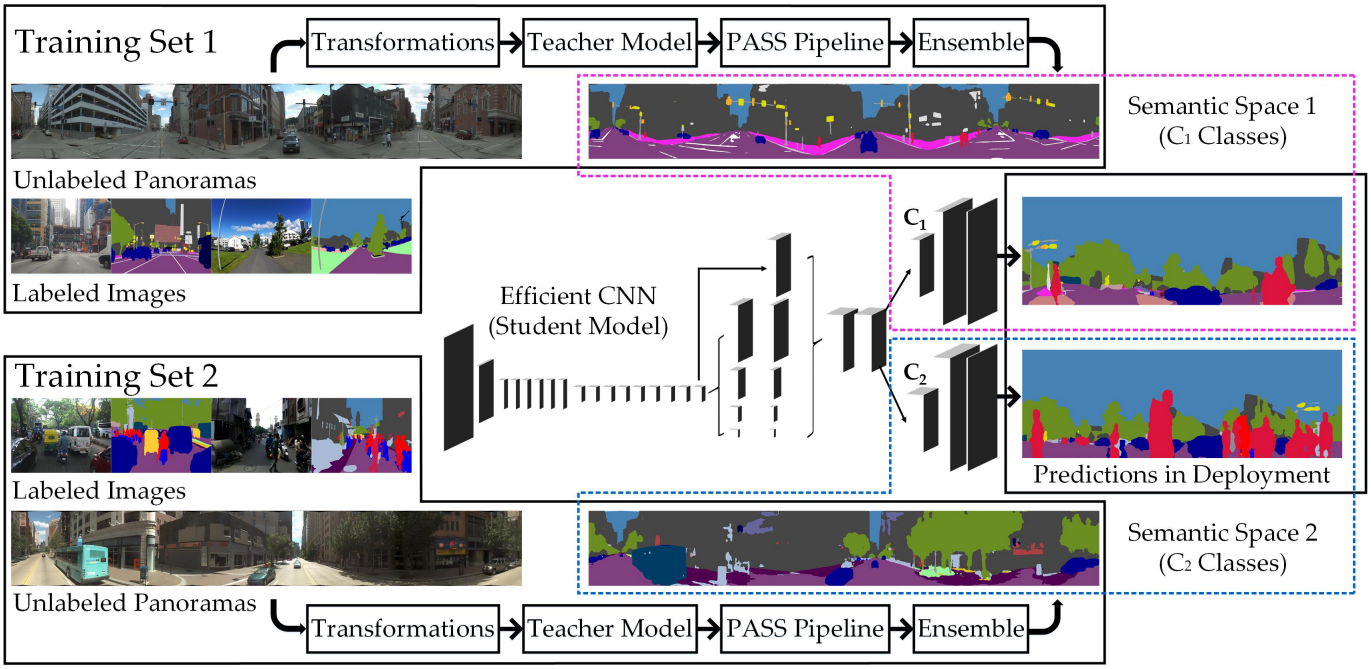


Fig. 2. Diagram of the proposed omnisupervised learning framework for omnidirectional semantic segmentation. During training, both labeled images and unlabeled panoramas are incorporated. Annotations for unlabeled panoramas are automatically generated by using the specialized ensemble of a teacher model's predictions on multiple transformations with the PASS pipeline. During deployment, the yielded student model is not only efficient and suitable for panoramas, but also robust and capable of delivering multiple sets of visual classes, enriching detectable semantics to fully understand unconstrained surroundings.

in the training stage. Although large-scale annotated omnidirectional vision datasets are scarce, there are a large amount of unlabeled panoramic images or videos available in the community. In this work, we leverage a source of unlabeled panoramas, and propose a method to automatically create their labels. Following the concept of data distillation [28], we use a large teacher model whose architecture may be sophisticated which disqualifies its usage in real-time applications. However, the teacher's produced segmentation maps are finely grained which provide the potential for data distillation. As it is depicted in Fig. 3, we ensemble the teacher model's predictions on various transformed copies of a panoramic image to produce the final annotation. Specifically, the ensemble process takes into account the wide-angle and wrap-around features of panoramas.

The teacher network is trained on conventional pinhole images. The teacher model F_t, can be 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. As it is shown in Fig. 3, each panorama segment I_i (size: $H_p \times \frac{W_p}{N}$) is fed into a feature model to predict a feature map. This helps to leverage the correspondence between the features inferred from a panorama segment and the features learned from pinhole images [16], as they both correspond to a similar narrow FoV, formally:

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

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

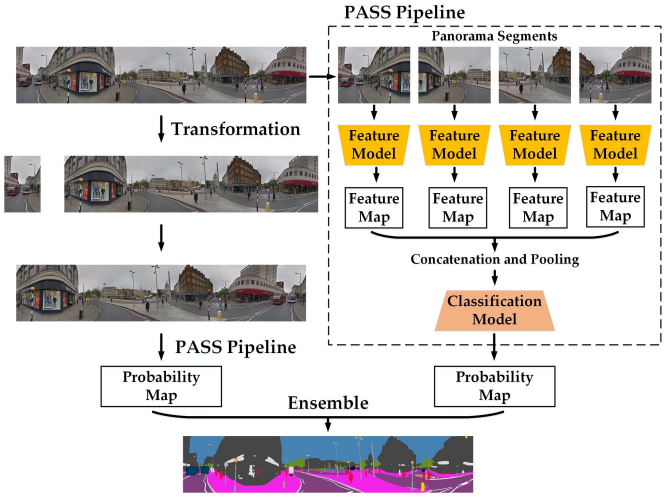
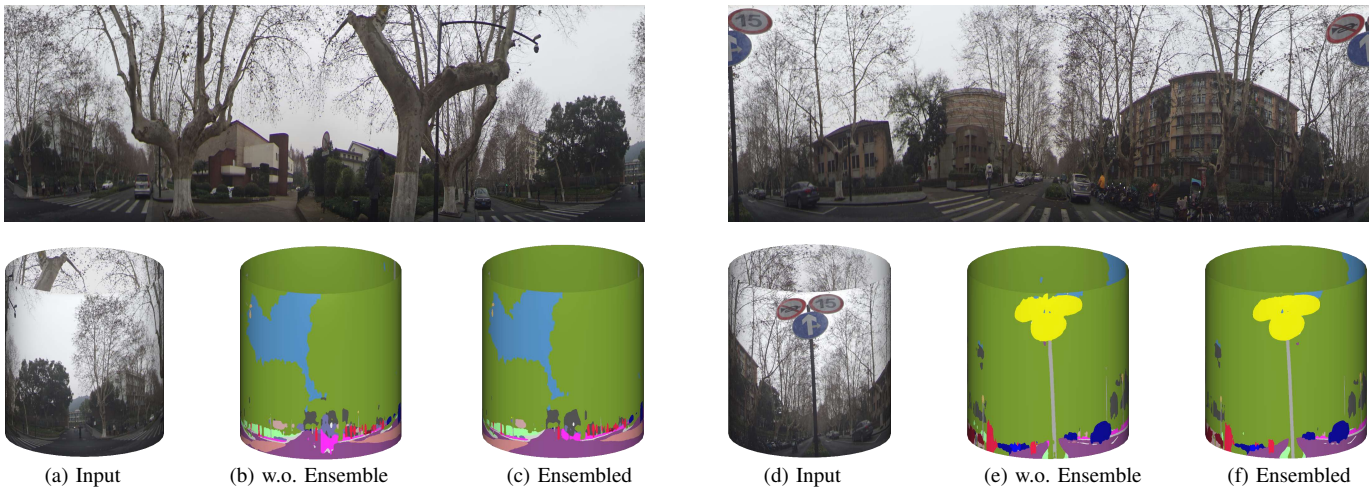


Fig. 3. Diagram of the specialized ensemble method with the PASS pipeline by considering the wide-angle and wrap-around connections of panoramas.

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

$$P_p^{H_p \times W_p} = F_{t_c}\left[\biguplus_{i=1}^{N} F_{t_e}\left(I_i^{H_p \times \frac{W_p}{N}}\right)\right] \tag{2}$$

This is due to that the classification model with lean convolution layers, which is also known as the fusion model in the PASS pipeline, is mainly responsible for the classification since the semantically-meaningful feature map has already



(a) Input (b) w.o. Ensemble (c) Ensembled (d) Input (e) w.o. Ensemble (f) Ensembled

Fig. 4. Ensemble helps to generate seamless semantic maps for data distillation as panoramas can be folded into 360° cylindrical rings.

been predicted and aggregated. As in many semantic segmentation networks with asymmetric architectures [12][13], it is preferred to have a powerful backbone for recognizing semantics and a light-weight upsampling path (e.g., 128 or 256 dimensions) for delineating borders. The fusion model has to be no leaner to take the local context around the borders of panorama segments into consideration [14], which is beneficial for continuous and smooth segmentation. PASS pipeline incurs nearly N times of computation of a single model, as the capacity mostly lies in the feature model. Thus, it is suitable to be used in the preparation stage other than deployment.

In addition, we propose a specialized ensemble method by taking into considerations the omnidirectional trait of our task and the wrap-around structure of panoramic images in the unfolding direction. Precisely, this is achieved by unfolding a panorama $I_p^{H_p \times W_p}$ from M evenly spaced positions or rotating the panorama for M times. 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. For instance, Fig. 3 depicts a rotation of 45°. Then, an ensemble of the predictions can be created, to have the final annotation A_u for an unlabeled panorama, formally:

$$A_u = \bigsqcup_{k=1}^{M} P_{p_k}^{H_p \times W_p} = \bigsqcup_{k=1}^{M} F_{t_c} \left[\biguplus_{i=1}^{N} F_{t_e} \left(I_{i_k}^{H_p \times \frac{W_p}{N}} \right) \right] \quad (3)$$

where $\bigsqcup$ denotes the ensemble process that can be achieved by averaging, weighing or aggregating the CNN's per-pixel probability maps for the transformed panoramas. An ensemble of predictions is more robust in nature (compared to a single pass), since averaging the knowledge of multiple predictions from one teacher network enables a model to be more prepared against unseen data. Although this has a direct negative impact on efficiency since making M predictions is always more complex than computing a forward pass, these operations (including PASS pipeline) are processed off-line in the preparation stage, and they help to better recover the decoupled contextual information. Overall, the proposed ensemble method helps to yield dense and more accurate segmentation maps that can be better trusted for data distillation. Qualitative examples of the produced fully seamless semantic maps without any blind zones are shown in Fig. 4.

B. Training Stage

To yield a segmentation model suitable for panoramic images, we propose to leverage multiple data sources, as a single training set is limited in the diversity of FoVs, which incurs a comparatively large overfitting risk, due to all images being collected with the same camera or certain types of acquisition setups [10]. Formally, 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}, with a semantic class space C_i. To train an efficient student CNN F_s, the conventional strategy is to learn the mapping represented by the following equation:

$$F_s \left(S_{il} \right) \Longrightarrow A_{il}(C_i) \quad (4)$$

The efficient segmentation model F_s, likewise, can also 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:

$$F_s \left(S_{il} \right) = F_{s_c} \left[F_{s_e} \left(S_{il} \right) \right] \Longrightarrow A_{il}(C_i) \quad (5)$$

The aim of our multi-source learning is to train a single model simultaneously in different domains, but the semantic spaces in disparate datasets are incompatible [64][65]. 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. For example, road surfaces are simply defined as road and sidewalk in Cityscapes [10] and IDD20K [15], but in Mapillary Vistas [11], they are labeled as road, sidewalk with curbs between them, and additional roadway classes like crosswalks. In addition, riders in IDD20K would be distinguished into motorcyclists and bicyclists if

from Vistas. There are novel safety-critical classes like auto-rickshaws that are absent in Cityscapes (European urban areas) but widespread in the unstructured IDD dataset (Asia).

In spite of the different class definitions, we argue that the relationships encoded in the similar label hierarchies could positively reinforce the generalizability of feature representations when learning with multiple disparate domains. For illustration, street-scene datasets both have flat (road, sidewalk), vehicle (car, bus), road-side object (curb, pole) classes, even though they are defined with inconsistent taxonomies. Therefore, we consider that it is fruitful to bridge multiple datasets for training our student model.

In Fig. 2, we illustrate our framework in the case of two training domains, but it can be easily scaled up to multiple domains. Precisely, to address the heterogeneity in the semantic spaces, we append two heads (classification models $F_{s_{c1}}$ and $F_{s_{c2}}$) to the efficient CNN architecture as depicted in Fig. 2, each of which is a fully convolutional module with an upsampling layer for prediction in the specific label space. Thereby, the training target can be modified into:

$$F_{s_{c1}}\left[F_{s_e}\left(S_{1l}\right)\right], F_{s_{c2}}\left[F_{s_e}\left(S_{2l}\right)\right] \Longrightarrow A_{1l}, A_{2l} \quad (6)$$

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

$$F_{s_{ci}}\left[F_{s_e}\left(S_{il}, S_u\right)\right] \Longrightarrow \left(A_{il}, A_{iu}\right)(C_i) \quad (7)$$

In this regard, the panoramas also serve as a domain bridge when performing joint training, as they will be fed to the student in both semantic spaces but not necessarily in the same forward/backward passes, which helps to yield more robust feature representations irrelevant of imagery domains.

C. Deployment Stage

After training, the student CNN is ready for being applied in omnidirectional imagery, while neither ensembling, fusing nor post-processing is needed in the deployment phase. The resulted student is a single model, which maintains the efficiency and simplicity as in the common case of semantic perception systems. Meanwhile, it possesses several important benefits. First, since the student has been exposed to omndirectional and heterogeneous data, its generalizability has been significantly enhanced in new panoramic domains. Second, the model is able to deliver diverse sets of detectable semantics:

$$\bigvee_{i=1}^{T} F_{s_{ci}}\left[F_{s_e}\left(I_{p_n}\right)\right] = \bigvee_{i=1}^{T}\left(P_i(C_i)\right) \quad (8)$$

where $\bigvee$ denotes a union of operations or semantic maps. For a new panoramic image I_{p_n}, T predictions of semantic maps will be generated, each of which P_i corresponds to a semantic space C_i, supposing a very rich resource of mutually complementary information for upper-level navigational applications. Overall, the omnisupervised solution enables the efficient CNN to become more prepared and suitable in any target panoramic domain to fully understand real-world unconstrained surroundings.

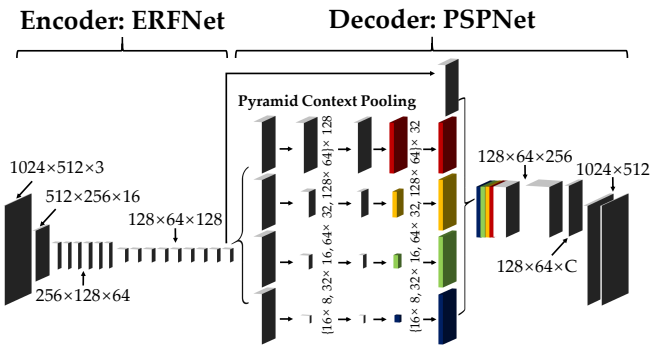


Fig. 5. Efficient semantic segmentation CNN architecture ERF-PSPNet with pyramid context pooling.

D. Student Learner Architecture

As illustrated in Fig. 5, the student learner follows the encoder-decoder sequential architecture of ERF-PSPNet [6], which was designed for navigation assistance applications. It leverages the encoder of the well-known ERFNet [12] to achieve a good trade-off between inference speed and segmentation accuracy. The encoder is attached with the pyramid pooling module in PSPNet [31], where the feature pyramid is upsampled and concatenated with the input features. Thus, subsequent convolutions obtain access to broad spatial pools which increase their receptive field. In this work, with the purpose of leveraging non-local strategies that help to capture rich global context-aware features available in omnidirectional images, we present two variants of the efficient ERF-PSPNet architecture.

As depicted in Fig. 5, the variants share a similar setup, both performing pyramidal context pooling in the decoder, which are denoted using different colors for different pyramid levels. Our critical modification lies in the combination of light-weight attention modules with the pyramid structure, to materialize the potential of global contextual information in full-view panoramas. Note that this has not been properly addressed in prior works [14][16] with the input panorama partitioned into pieces for several forward passes. Contrarily, our omnisupervised solution that covers panoramic imagery in the training, helps to unlock the use of long-range contextual dependencies.

Precisely, the first variant is ERF-PSPNet+OC (Object Context), where the pyramidal pooling module has been appended with object context aggregation in each pyramid scale. Specifically, OC denotes Object Context, whose estimation [35] is to produce a fully dense affinity matrix that measures the similarities of each pixel and each other pixel for the whole feature map. In this way, ERF-PSPNet+OC can exploit the degree that indicates pixels fall in the same semantic class, by transforming the feature map into a per-pixel weighted one according to the similarities. The second variant is ERF-PSPNet+scSE (spatial and channel Squeeze-Excitation), where

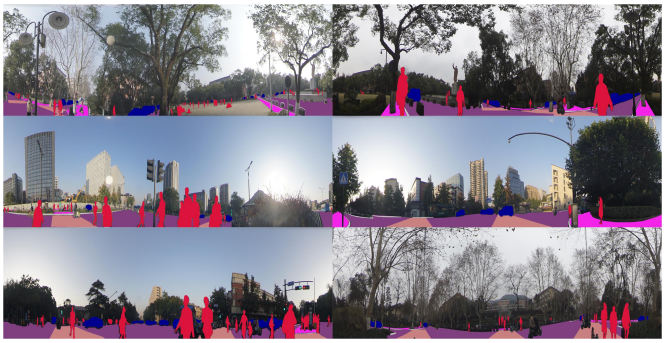


Fig. 6. Examples from the extended PASS dataset with annotations on navigation-critical classes: Car, Road, Sidewalk, Crosswalk, Curb and Person.

the pyramidal pooling module has been appended with concurrent spatial and channel attention [66], which separately recalibrates the feature maps to be more informative along channel and space. Here, scSE denotes spatial and channel Squeeze-Excitation, where the concurrent recalibration extends the squeeze-and-excitation attention [46] in a way that complementarily squeezes channel-wise and excites spatially. The scSE component relaxes the local context constraint and provides more importance to relevant spatial locations, which are meaningful features to be exploited in panoramas. In this sense, it can also be considered as a non-local module when embedded in our ERF-PSPNet+scSE structure with a feature pyramid. For both variants, we perform the context aggregation at low resolutions before upsampling and concatenation with the original features from the encoder. Thereby, such seamless integration comes with a slight increase in complexity.

IV. Experiments

A. Datasets

Target Testing Dataset. We target the Panoramic Annular Semantic Segmentation (PASS) database [16], a challenging dataset to assess the robustness and real-world applicability of panoramic semantic perception algorithms. It is intended for testing the generalizability of models trained on other datasets and does not provide a training set of its own. Representing an unseen omnidirectional domain, it is a testbed to investigate the benefits of our omnisupervised solution and compare the yielded student model with known efficient CNNs. The PASS dataset contains 400 raw annular and annotated unfolded panorama pairs. As an evaluation dataset, the original version [16] has annotations on 4 classes. In this work, we update the dataset by creating pixel-accurate annotations on 6 navigation-critical classes: Car, Road, Sidewalk, Crosswalk, Curb and Person (see Fig. 6).

Multi-Source Training Datasets. Our multi-source training is experimented with two conventional pinhole image datasets: Mapillary Vistas [11] and IDD20K [15], two of the richest street scene parsing datasets nowadays. Vistas offers images with high diversity shot by various cameras across continents. In addition, it covers a variety of viewpoints with data captured from the perspective of vehicles (roadways) and pedestrians (sidewalks). This variability is especially appealing for omnidirectional semantic segmentation because it exposes the learner to a wide array of observations other than only forward-facing views. Regarding IDD20K, it imports extremely unstructured environments, which is attractive due to the cluttered scenes implicated in panoramic imagery.

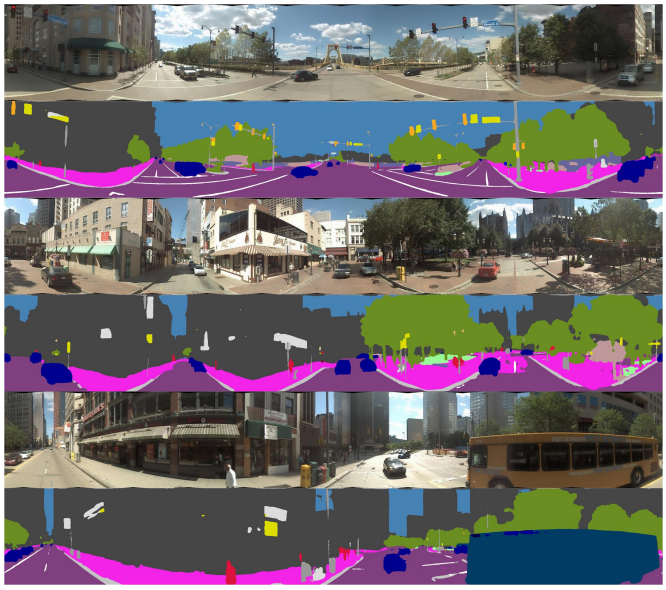


Fig. 7. Examples of automatically generated panoramic annotations for the stitched panoramas using Pittsburgh dataset [67].

Vistas contains 18000/2000/5000 images for training, validation and testing. IDD20K comprises 14027/2036/4038 images in the train/val/test subsets. Both ground-truth labels for the testing images are not openly available, but in this work the PASS dataset is readily accessible for evaluation. For Vistas, we use the 25 classes for training and report the segmentation accuracy on the validation set as shown in Table I. For IDD20K, we report results for the level-3 labels (26 classes) on the validation set, as shown in Table II. We adopt the standard Intersection-over-Union (IoU) as the evaluation metric:

$$IoU = \frac{TP}{TP + FP + FN} \tag{9}$$

where TP, FP and FN are respectively the number of true positives, false positives and false negatives at pixel level.

Unlabeled Panoramas Dataset. Regarding the unlabeled panoramas for data distillation, we use the Pittsburgh dataset [67]. 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. We stitch the lower pitch images whose perspective matches road-driving imagery by using the known stitching method [68]. Overall, we obtain 966 stitched panoramas from the query set. Fig. 7 displays examples of the stitched panoramas and 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. It should also be noted that the panoramas provide rich and distinctly different global contextual cues from those of conventional pinhole images, e.g., various directions of roadways and sidewalks can be simultaneously observed.

TABLE I
CLASS-WISE SEGMENTATION ACCURACY OF THE JOINTLY-TRAINED ERF-PSPNET ON MAPILLARY VISTAS DATASET [11]. **POL**, **STL**, **BIL** ETC. ARE ABBREVIATIONS OF THE CLASSES: POLE, STREET LIGHT, BILLBOARD, TRAFFIC LIGHT, CAR, TRUCK, BICYCLE, MOTORCYCLE, BUS, TRAFFIC SIGN FRONT, TRAFFIC SIGN BACK, ROAD, SIDEWALK, CURB, FENCE, WALL, BUILDING, PERSON, MOTORCYCLIST, BICYCLIST, SKY, VEGETATION, TERRAIN, ROAD MARKING AND CROSSWALK. **MEAN IOU (MIOU)**: 63.0%.

Pol	**StL**	**Bil**	**TrL**	**Car**	**Tru**	**Bic**	**Mot**	**Bus**	**SiF**	**SiB**	**Roa**	**Sid**
49.6%	27.5%	42.9%	56.9%	90.7%	65.6%	55.3%	55.1%	74.4%	66.9%	29.0%	90.8%	70.3%
Cur	**Fen**	**Wal**	**Bui**	**Per**	**MoC**	**BiC**	**Sky**	**Veg**	**Ter**	**Mar**	**Cro**	**mIoU**
58.2%	55.9%	51.4%	86.5%	71.9%	54.0%	49.3%	98.2%	89.9%	67.1%	53.7%	64.6%	63.0%

TABLE II
CLASS-WISE SEGMENTATION ACCURACY OF THE JOINTLY-TRAINED ERF-PSPNET ON IDD20K DATASET [15]. **ROA**, **DRF**, **SID** ETC. ARE ABBREVIATIONS OF THE CLASSES: ROAD, DRIVABLE FALLBACK, SIDEWALK, NONDRIVABLE FALLBACK, PEDESTRIAN, RIDER, MOTORCYCLE, BICYCLE, AUTO RICKSHAW, CAR, TRUCK, BUS, VEHICLE FALLBACK, CURB, WALL, FENCE, GUARD RAIL, BILLBOARD, TRAFFIC SIGN, TRAFFIC LIGHT, POLE, OBSTACLE FALLBACK, BUILDING, BRIDGE, VEGETATION AND SKY. **MEAN IOU (MIOU)**: 64.2%.

Roa	**DrF**	**Sid**	**NoF**	**Ped**	**Rid**	**Mot**	**Bic**	**AuR**	**Car**	**Tru**	**Bus**	**VeF**
93.3%	63.0%	66.2%	49.7%	65.1%	69.1%	73.7%	34.8%	83.0%	88.0%	79.7%	86.7%	41.5%
Cur	**Wal**	**Fen**	**GuR**	**Bil**	**TrS**	**TrL**	**Pol**	**ObF**	**Bui**	**Bri**	**Veg**	**Sky**
74.2%	56.2%	41.1%	50.8%	60.2%	56.4%	23.2%	48.0%	42.8%	72.1%	65.7%	87.4%	96.7%

(a) Panorama with ground-truth annotation (b) Prediction without any ensemble (c) Prediction with our ensemble method

Fig. 8. Qualitative comparison of the predictions by the teacher model PSPNet50 [31]: (a) Panoramas with ground-truth annotations, (b) Predictions of PSPNet50 without any ensemble nor PASS pipeline, (c) Predictions of PSPNet50 with the proposed ensemble method.

B. Training Setups

Teacher Network. We use PSPNet50 [31] as the teacher model, which has mean IoU (mIoU) of 67.1% on Vistas and 66.5% on IDD20K. We experiment with different ensemble methods and their combinations to generate annotations for the panoramas. As shown in Table III, PSPNet achieves 41.4% in the panoramic domain when testing without the PASS pipeline, and achieves 70.6% with the PASS pipeline but without any ensemble. This demonstrates that although the PASS pipeline incurs more computation, the proposed usage in the data preparation stage helps to generate significantly more accurate panorama annotations. Regarding ensemble methods, multi-scale prediction is widely used in semantic segmentation methods [8][32], which helps to attain higher accuracy. Horizontal flipping (mirroring) also helps to improve the overall prediction certainty. Since the PASS pipeline separates the panorama into 4 segments [16], the specialized ensemble method starts with at least 8 times of rotations. As shown in Table III, the specialized ensemble strategy consistently boosts the mIoU, which becomes saturated until 32 times of rotations. Thereby, we use ensemble-32, and generate duplicates of panoramas jointly with multi-scale and mirroring transformations. We ensemble the teacher model's predictions by aggregating the probability maps for these copies to form as the annotation of the unlabeled panorama set for data distillation. Although ensemble increases the time cost, it is automatically conducted and only used in the data preparation phase, which does not impair the efficiency of the student model. Finally, the accuracy boosts to 71.9%. Fig. 8 shows the predictions of PSPNet50 on panoramas from PASS dataset, which demonstrates that overall the ensemble method significantly improves the quality of the semantic maps. With the certainty and continuity benefits that are also demonstrated in Fig. 4, the automatically generated panoramic annotations are more trusted for data distillation.

Student Network. For the student architecture, we experiment with ERF-PSPNet [6], due to its real-time performance, publicly available ImageNet [69] pre-trained weights

TABLE III
ACCURACY ANALYSIS OF ENSEMBLE METHODS FOR VISTAS-TRAINED TEACHER MODEL PSPNET50 [31] ON PASS DATASET. ENSEMBLE-M: THE PANORAMA IS ROTATED FOR M TIMES WITH VARIATION OF $360°/M$ TO FORM AS M TRANSFORMATIONS.

Ensemble method	mIoU
Without PASS	41.4%
Without ensemble	70.6%
Multi-scale (MS)	71.3%
Mirror (MI)	71.1%
Ensemble-8	71.2%
Ensemble-16	71.3%
Ensemble-32	71.4%
Ensemble-64	71.4%
Ensemble-32+MS+MI	**71.9%**

TABLE IV
ACCURACY ANALYSIS ON IDD20K DATASET [15] IN MIOU.

Network	IDD20K	Vistas
SegNet [29]	38.4%	NA
DRNet (ResNet18) [30]	52.2%	NA
USSS (ResNet18) [65]	27.5%	NA
USSS (ResNet50) [65]	55.1%	NA
ERFNet [12]	55.4%	NA
ERF-PSPNet (Vistas-trained)	NA	61.6%
ERF-PSPNet (IDD-trained)	63.2%	NA
ERF-PSPNet (Jointly-trained)	**64.2%**	63.0%
ERF-PSPNet (Omnisupervised)	64.0%	62.9%
ERF-PSPNet+OC (Omnisupervised)	**64.2%**	63.2%
ERF-PSPNet+scSE (Omnisupervised)	**64.2%**	**63.6%**

and capability to exploit rich contextual priors. The models are trained under Adam optimization [70] with a Weight Decay of 2.0×10^{-4} and an initial Learning Rate of 5.0×10^{-4} that decreases exponentially over 200 epochs. We feed the samples with a batch size of 12 and a resolution of 1024×512 as a balance between the two heterogeneous training sets (Vistas+IDD20K). For the multi-source training, each iteration comprises a forward pass and a backward pass per dataset using cross-entropy loss functions. In the omnisupervised setting, we also experiment with the ERF-PSPNet variants to study whether they help to gather more global context-aware features in the panoramas.

C. Semantic Segmentation Accuracy

As shown in Table IV, the ERF-PSPNet trained independently on Vistas and IDD20K achieves 61.6% and 63.2% of mIoU on the respective validation dataset. The joint-training boosts the scores to 63.0% and 64.2%, which demonstrates the benefit of multi-source supervision. The improvement is due to the more generalized feature representation thanks to our framework that bridges multiple datasets in training. These results surpass those of previous efficient networks including SegNet [29], DRNet [30] and ERFNet [12] attempted on IDD dataset. Compared with the USSS approach [65] which also trains on multiple sources through a semi-supervised solution, our score is significantly higher because we are able to leverage the full IDD20K and Vistas supervision. In Table I and Table II, we present the class-wise accuracies on Vistas and IDD20K validation sets obtained by our multi-source training. This sets the new state of the art among efficient semantic segmentation CNNs on the challenging unstructured IDD20K database. Regarding the omnisupervised solution, it slightly decreases the accuracy on the validation sets, which is reasonable as the validation data only contain pinhole images, but the student becomes more generalized and ready to be deployed in open omnidirectional imagery.

D. Generalization in Panoramic Domain

We step further to study the generalizability in panoramic domain by using the Panoramic Annular Semantic Segmentation (PASS) database. In Table V, we present the accuracy, Memory Access Costs (MACs) and Parameters (PARAMs) of four computation-expensive networks SegNet [29], PSPNet50 [31], DenseASPP [33] with DenseNet121 [9] and DANet [34] with ResNet50 [1], both trained on Mapillary Vistas by using the proposed hyper-parameters for them in their respective publications. They are tested by viewing the panorama as a single segment without any ensemble nor separation. We also include a variety of efficient networks including ENet [38], LinkNet [39], SwiftNet [13], SwaftNet [14], etc. These real-time networks experimented by [14], are also tested in an end-to-end way. But they rely on a heterogeneous set of data augmentation and style transfer-based domain adaptation strategies, which are known beneficial for improving the performance in target domains [16][25]. In comparison, our approach is more realistic as normally the style and knowledge about the target domain are inaccessible. Overall, as shown in Table V, our omnisupervised solution with ERF-PSPNet outperforms state-of-the-art networks on PASS dataset without using any domain adaptation strategy nor increasing any computation complexity.

On the Benefit of Multi-Source Training. PASS is a highly unconstrained domain as in panoramic imagery, traffic participants with diverse orientations can be simultaneously observed (see examples in Fig. 2). In addition, sometimes there are many close pedestrians present in the images from PASS dataset. In Table VI, we perform ablation studies and analyses of different training methods. As the annotations of the PASS dataset for testing were created according to the labels definition of Mapillary Vistas, in most cases we evaluate the performance with the Vistas-space result. However, we could also evaluate by using the IDD-space result, but the score is lower due to the discrepancy of the classes. For instance, in IDD20K space, curb class is defined in a different way. Accordingly, as shown in Table VI, Vistas-trained and IDD-supervised ERF-PSPNet achieves 32.2% and 20.1% of mIoU on the PASS dataset, respectively.

Surprisingly, our jointly-supervised ERF-PSPNet based on Vistas and IDD20K boosts the mIoU to 41.0%, significantly higher than independent-training results (20.1% and 32.0%). The huge robustness gains are obtained owing to the well generalized feature representation offered by our training with multiple heterogeneous datasets. Specifically, IDD20K-supervision exhibits the student to highly unstructured scenes while Vistas-supervision affords the high data diversity, which

TABLE V
ACCURACY AND COMPUTATION COMPLEXITY ANALYSIS ON PANORAMIC ANNULAR SEMANTIC SEGMENTATION (PASS) DATASET. ALL NETWORKS ARE TESTED BY VIEWING THE PANORAMA AS A SINGLE SEGMENT WITHOUT ANY SEPARATION.

Network	Car	Road	Sidewalk	Crosswalk	Curb	Person	mIoU	MACs	PARAMs
SegNet [29]	57.5%	52.6%	17.9%	11.3%	11.6%	3.5%	25.7%	398.3G	28.4M
PSPNet (ResNet50) [31]	76.2%	67.9%	34.7%	19.7%	27.3%	22.6%	41.4%	403.0G	53.3M
DenseASPP (DenseNet121) [33]	65.8%	62.9%	30.5%	8.7%	23.0%	8.7%	33.3%	78.3G	8.3M
DANet (ResNet50) [34]	70.0%	67.8%	35.9%	21.3%	12.6%	25.9%	38.9%	114.1G	47.4M
ENet [38]	59.4%	59.6%	27.1%	16.3%	15.4%	8.2%	31.0%	4.9G	0.4M
LinkNet [39]	62.6%	64.9%	23.2%	6.6%	18.1%	7.5%	30.5%	25.8G	11.5M
SQNet [40]	56.5%	57.2%	19.1%	21.4%	10.4%	3.0%	27.9%	206.7G	16.3M
ICNet [41]	49.3%	52.4%	20.0%	16.7%	6.7%	9.3%	25.7%	10.3G	11.6M
ESPNet [42]	52.6%	51.4%	21.6%	10.5%	6.5%	5.6%	24.7%	5.6G	0.2M
EDANet [43]	61.4%	64.0%	28.1%	6.3%	15.2%	8.1%	30.5%	9.0G	0.7M
BiSeNet [44]	61.8%	58.3%	17.3%	12.7%	10.8%	5.3%	27.7%	26.1G	12.9M
CGNet [45]	65.2%	56.9%	23.7%	3.8%	11.2%	21.4%	30.4%	7.0G	0.5M
ERFNet [12]	70.0%	57.3%	25.4%	22.9%	15.8%	15.3%	34.3%	30.3G	2.1M
PSPNet (ResNet18) [31]	64.1%	67.7%	31.2%	15.1%	17.5%	12.8%	34.8%	235.0G	17.5M
ERF-PSPNet [7]	71.8%	65.7%	32.9%	29.2%	19.7%	15.8%	39.2%	26.6G	2.5M
ERF-APSPNet [16]	72.3%	71.4%	32.6%	5.6%	16.3%	14.5%	35.5%	26.6G	2.5M
SwiftNet [13]	67.5%	70.0%	30.0%	21.4%	21.9%	13.7%	37.4%	41.7G	11.8M
SwaftNet [14]	76.4%	64.1%	33.8%	9.6%	26.9%	18.5%	38.2%	41.8G	11.9M
ERF-PSPNet (Omnisupervised)	**81.4%**	**71.9%**	**39.1%**	**24.6%**	**26.4%**	**44.1%**	**47.9%**	26.6G	2.5M
ERF-PSPNet+OC (Omnisupervised)	85.1%	76.8%	41.2%	11.8%	27.9%	53.9%	49.5%	32.5G	2.5M
ERF-PSPNet+scSE (Omnisupervised)	**83.3%**	**75.4%**	**46.8%**	**33.3%**	**28.2%**	**51.3%**	**53.0%**	26.7G	2.6M

are both important for robust semantic segmentation in omnidirectional imagery. Overall, our multi-source joint-training already achieves greater mIoU than the best efficient networks previously attempted on PASS dataset.

On the Benefit of Omni-Supervised Training. Moreover, our multi-source omnisupervised solution, denoted as ERF-PSPNet (Omnisupervised) in Table VI, further dramatically boosts the mIoU from 41.0% to 47.9%, even exceeding by 6.5% and 9.7% the computation-intensive PSPNet and the detail-sensitive SwaftNet, respectively. This increase is due to our omnisupervised proposal that exposes the student to omnidirectional data in the training phase.

We additionally compare with two training methods. The first is to directly train the model using Pittsburgh dataset (panoramic data) with the generated labels, which achieves 29.7%. The second is to finetune the model ERF-PSPNet (IDD20K+Vistas) on the Pittsburgh dataset also in a multi-source training way, which decreases the accuracy from 41.0% to 39.0%. They are not helpful due to the low diversity in the training panoramas and the gap between training/testing panoramic domains. In contrast, our omnisupervised solution effectively bridges multiple sources, successfully exploiting both the diversity in pinhole images and the knowledge in panoramic data, therefore the mIoU is rocked to 47.9% in the unseen omnidirectional domain.

We also compare with other distillation methods. First, we experiment with standard data distillation methods that use conventional pinhole images instead of panoramic images. We leverage the unlabeled images in the testing sets of Mapillary Vistas and IDD20K. It achieves 38.0% when trained in a multi-source manner. In another setting, we perform data distillation with Mapillary Vistas and ADE20K [71], which achieves 37.8%, still far lower than our approach. These show that the benefit of our omnisupervised solution is not merely owing to the increased amount of training samples but also importantly the knowledge in omnidirectional images. Second, we compare with two distillation methods including Local knowledge distillation [58] by distilling the probability outputs, and Relational knowledge distillation [57] by using the distance-wise loss [57] on the feature maps before final upsampling and classification. They are both trained in the multi-source manner. However, both of them are not as effective as our omnisupervised approach in terms of mIoU. Finally, we evaluate in the single-source data distillation setting by using the panoramas with automatically generated labels in the Vistas space, which achieves 46.3%, clearly lower than our multi-source omnisupervised solution.

To further verify that the proposal allows the student to exploit the global contextual information in the panoramas, we perform experiments with the architecture variants ERF-PSPNet+OC and ERF-PSPNet+scSE. Both the non-local attention operations employed in these architectures help to capture global context cues [35][66]. Thus, the omnisupervised solution based on these architecture variants leads to higher mIoU (49.5% and 53.0%). The benefits are more evident on the panoramic dataset than those on the pinhole datasets (see Table IV). This indicates that the leveraged non-local operations are exceptionally beneficial for exploiting the global features that are distinctly rich and important in the panoramic data. It again demonstrates the significance of our omnisupervised solution that covers panoramic imagery in the training, and therefore enables the student to access and harvest the rich contextual priors in panoramas. Especially, ERF-PSPNet+scSE clearly stands out in front of previous works, outstripping by large margins the scores achieved with state-of-the-art accuracy/efficiency-oriented networks.

In Table VII, we compare our omnisupervised ERF-PSPNets against three high-efficiency networks as well as DenseASPP [33], a top-accuracy sophisticated architecture

TABLE VI
ACCURACY ANALYSIS OF DIFFERENT TRAINING AND DISTILLATION METHODS ON PANORAMIC ANNULAR SEMANTIC SEGMENTATION (PASS) DATASET. BOLD FONTS DENOTE THE METHODS OF THIS WORK.

Method	Car	Road	Sidewalk	Crosswalk	Curb	Person	mIoU
Vistas-trained	68.8%	62.0%	26.6%	3.9%	17.5%	14.1%	32.2%
IDD20K-trained	53.4%	51.2%	3.2%	0.0%	2.3%	10.6%	20.1%
Pittsburgh-trained	71.1%	68.7%	26.3%	0.0%	11.9%	0.2%	29.7%
Multi-source trained (IDD20K+Vistas)	**75.5%**	**70.9%**	**32.5%**	**13.0%**	**20.6%**	**33.5%**	**41.0%**
Fine-tuned (IDD20K+Vistas)	82.6%	74.1%	21.6%	0.6%	25.7%	29.4%	39.0%
Standard data distillation (IDD20K+Vistas)	74.4%	68.5%	31.3%	10.9%	18.3%	24.3%	38.0%
Standard data distillation (ADE20K+Vistas)	68.4%	74.0%	39.6%	11.6%	18.7%	14.8%	37.8%
Relational knowledge distillation [57]	81.4%	75.6%	35.8%	15.2%	28.0%	38.8%	45.8%
Local knowledge distillation [58]	81.9%	70.9%	26.6%	0.0%	26.9%	16.9%	37.2%
Single-source data distillation	83.9%	77.4%	46.4%	4.1%	25.4%	40.8%	46.3%
Our omnisupervised solution	**81.4%**	**71.9%**	**39.1%**	**24.6%**	**26.4%**	**44.1%**	**47.9%**

while only requiring moderate computation power. We compare the accuracy in mIoU on PASS dataset (the same in Table V) and speed in Frames Per Second (FPS) at the 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. We report the mean FPS values over 400 forward passes running through all panoramas in the PASS dataset. It can be seen that our ERF-PSPNets reach high accuracies while running with inference speeds far above the real-time constraint. In addition, the ERF-PSPNet variants with attention operations only increase the computation complexity by small fractions. A comprehensive speed comparison with more efficient networks can be found in [16].

TABLE VII
ACCURACY IN MIOU AND SPEED IN FRAMES PER SECOND (FPS) OF OMNISUPERVISED ERF-PSPNET VARIANTS COMPARED WITH THE STATE OF THE ART ON PASS DATASET [16].

Network	Accuracy	Speed on GPU processors		
	PASS	Titan RTX	1080Ti	2080Ti
DenseASPP [33]	33.3%	70.6	30.8	57.7
ERFNet [12]	34.3%	148.7	78.4	105.6
EDANet [43]	30.5%	152.7	76.9	106.4
CGNet [45]	30.4%	93.1	49.0	68.7
ERF-PSPNet	47.9%	**164.7**	**91.2**	**132.1**
ERF-PSPNet+OC	49.5%	121.3	72.0	87.6
ERF-PSPNet+scSE	**53.0%**	139.5	74.3	115.0

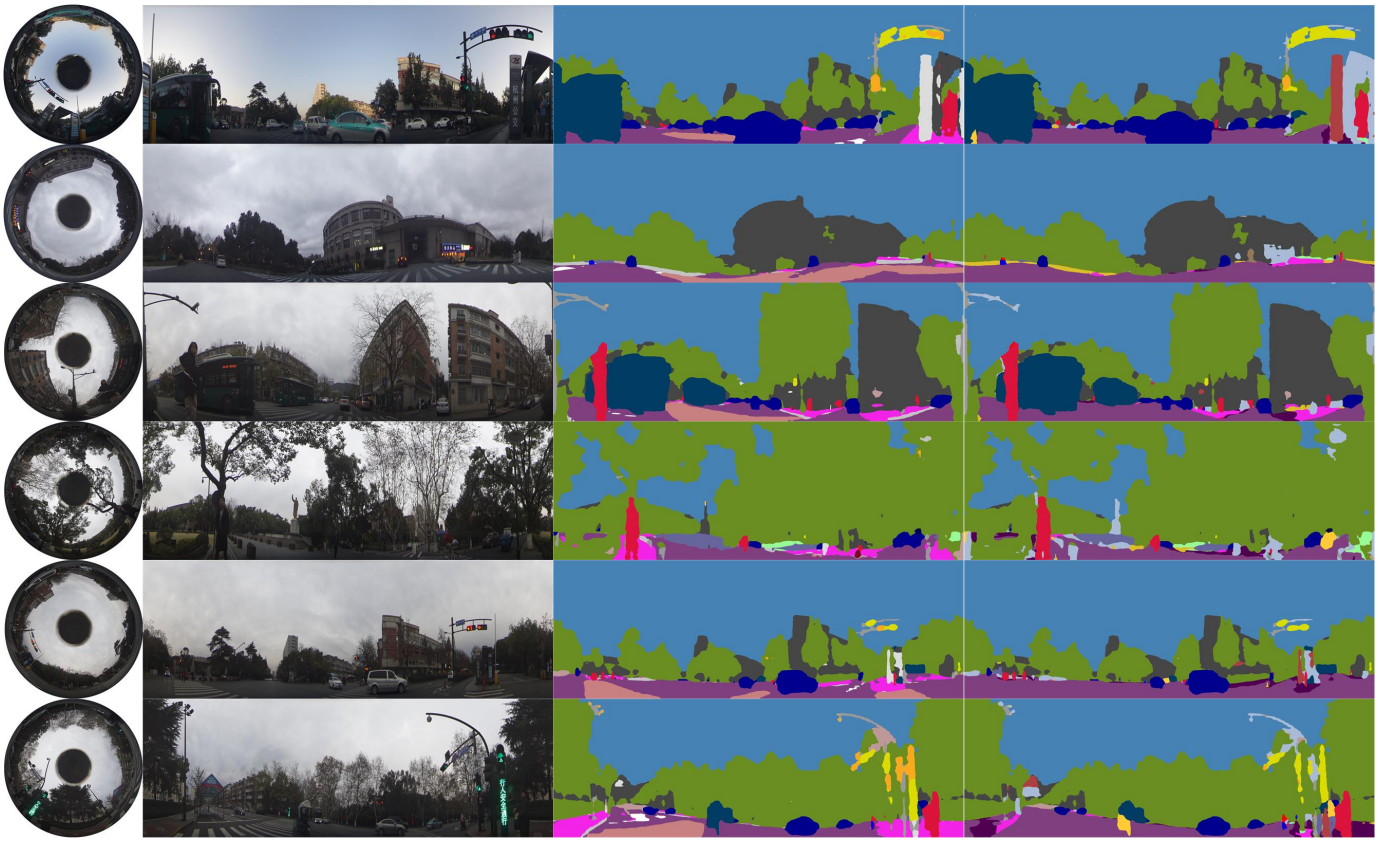
E. Qualitative Analysis and Discussion

Fig. 9 displays representative predictions of our omnisupervised ERF-PSPNet in multiple semantic spaces. On the one hand, clear and highly robust segmentation can be observed. 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 example, in the Vistas space, crosswalks and curbs can be predicted which are absent in IDD20K space (see Fig. 9), but IDD20K-space results can help to foresee safety-critical classes like auto-rickshaws whose behavior is highly unpredictable, as shown from the 4th to 6th rows (denoted in yellow). As a result, the detectable semantics and recognizable classes have been enriched, which are often required to fully understand real-world unconstrained surroundings. Fig. 10 further visualizes predictions of our approach in contrast with the state-of-the-art top-accuracy network DANet. It can be seen that DANet suffers from large accuracy downgrade when taken to omnidirectional imagery, while our omnisupervised high-efficiency ERF-PSPNet maintains a good performance in the previously unseen panoramic domain.

We also perform an experiment by testing with different FoVs of inputs in the panorama. As shown in Fig. 11, we crop different FoVs around the unfolded panorama center with variations of 10° for each point, and input the cropped images from the PASS dataset to the Vistas-trained ERF-PSPNet (baseline) and Omni-supervised ERF-PSPNet (our proposal). As it can be clearly seen, our proposal outperforms the baseline for all FoVs of inputs, while the gap is especially large for ultra-wide angles. In this sense, the proposal is not strictly tied to 360° panorama segmentation, but can be effective and deployed with other wide-angle omnidirectional cameras, e.g., fisheye, catadioptric and polydioptric lenses. However, while our omnisupervised solution has reaped huge generalization gains for large-FoV omnidirectional data, the performance is still not as accurate as the most comfortable angles. Thereby, if the computation budget is available, one can also use the PASS pipeline [16], or the ensemble methods in the deployment to further improve the reliability.

F. Deployment on an Instrumented Vehicle

To verify the applicability for IV systems, we deploy our omnisupervised ERF-PSPNet on a fully electric instrumented vehicle as shown in Fig. 12a. We use a newly designed panoramic annular lens system, which images a FoV of 360°×70°. It has been installed on the top of the instrumented vehicle with a LiDAR sensor. It has a FoV of 40° above the horizontal plane, and a surrounding view with 30° vertical FoV below, which allows to perceive more information about the roadways than the PASS data [16]. We use an NVIDIA Jetson Nano that is very portable to be deployed on driving systems, where a single forward pass of ERF-PSPNet achieves 15.9FPS at the resolution of 1024×512. In contrast, the previous PASS [16] system runs only at 4.2FPS due to the computation-intensive operations burdened on the segmenter.



(a) Raw (b) Unfolded (c) Prediction in Vistas space (d) Prediction in IDD20K space

Fig. 9. Qualitative examples of omnisupervised omnidirectional semantic segmentation on images from PASS dataset: (a) Raw, (b) Unfolded panoramas, (c) Predictions in Vistas space and (d) IDD20K space.

This demonstrates that our omnisupervised solution enables to fulfill omnidirectional semantic segmentation in near real time even on an embedded processor, because the efficient CNN is directly applicable via a single pass, without any panorama separation nor domain adaptation that would incur additional computationally costly processing.

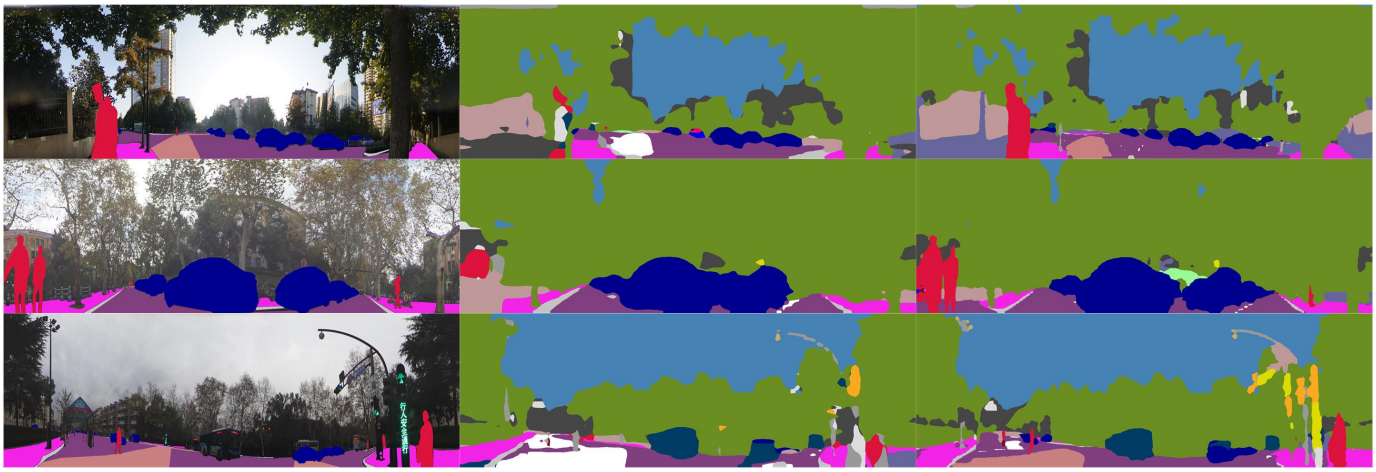
We perform a cross-season data collection with the instrumented vehicle at the Chengyuan campus of Zhejiang University in both summer and winter (see Fig. 12). This dataset has also been made publicly available together with the extended PASS dataset. Owing to the proposed solution that benefits the generalization and the panoramic imaging FoV that matches with the instrumented vehicle, the predicted semantic maps are highly precise and robust under different weather and illumination conditions.

Furthermore, we experiment by virtually navigating the instrumented vehicle in nine cities: New York, Beijing, Shanghai, Changsha, Hangzhou, Huddersfield, Madrid, Karlsruhe and Sydney, which exhibits a significantly larger variability in terms of geographic origins. This is achieved by using Google Street View to generate panoramas of the regions. For each panorama of 13312×6656 pixels, we crop 70° of vertical FoV with the pitch directions from -30° to 40°, which corresponds to the FoV of the panoramic annular camera on the instrumented vehicle. For each city, we gather 20 panoramas, to form as a set of 180 panoramas provided to the community. As shown in Fig. 13, the proposed omnisupervised solution helps the student generalize well to broader areas of the world than our campus thanks to multi-source data distillation. Even in adverse conditions such as the nighttime and rainy scenes (see Beijing and Shanghai examples in the 2nd/3rd rows of Fig. 13), panoramic segmentation maintains qualified. In summary, from both efficiency and generalizability perspectives, our approach is ideally suitable for IV applications. Meanwhile, it leaves rich opportunities to fuse the 360° semantics with LiDAR point clouds for complete scene understanding.

V. Conclusions

In this work, we have proposed an omnisupervised omnidirectional semantic segmentation framework, which bridges multiple heterogeneous data resources. We show that omnisupervised learning approached by multi-source data distillation has empowered the learner to gain significant benefits, making efficient CNNs like our ERF-PSPNet become suitable for wide FoV omnidirectional semantic segmentation. While only a single model is yielded in the omnisupervised training, it delivers multiple sets of visual classes, enriching the detectable semantics required to fully understand complex real-world unconstrained surroundings.

We show that 360° scene parsing can be addressed by running the single model in a single pass, without any panorama separation, fusion nor domain adaptation, so the inference efficiency of the CNN is maintained in the deployment phase.



(a) Panorama with ground-truth annotation (b) DANet50 (c) Our omnisupervised ERF-PSPNet

Fig. 10. Qualitative comparison of our approach with the state of the art: (a) Panoramas with ground-truth annotation, (b) Predictions of DANet (with ResNet50) [34], (c) Predictions of our solution (omnisupervised ERF-PSPNet).

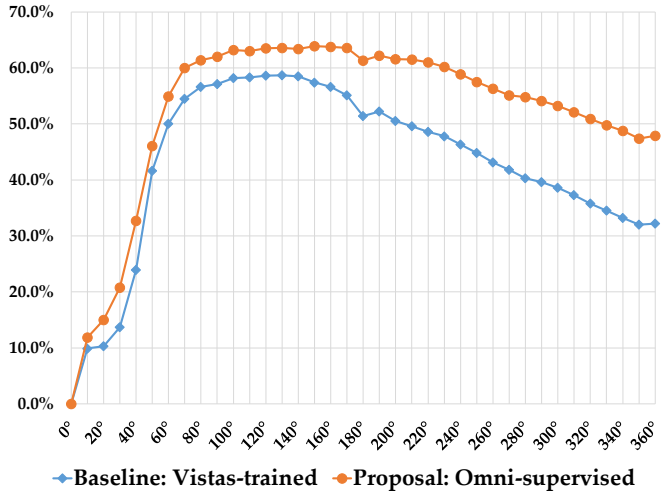


Fig. 11. Accuracy curves of the proposed omni-supervised ERF-PSPNet compared with the baseline measured in mIoU by using different FoVs of inputs from the panoramas.

Besides, it allows the segmenter to exploit the distinctly rich global context-aware features in the panoramas with non-local attention operations. We integrate the omnisupervised ERF-PSPNet on an instrumented vehicle with a panoramic annular lens, demonstrating that from both efficiency and generalizability perspectives, the system is ideally suitable for IV applications. In the future, we aim to obtain a large-scale panoramic dataset with images from all of the world, and achieve panoramic panoptic segmentation to enable a more unified scene perception.

REFERENCES

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

[2] W. Wang, J. Shen, and H. Ling, "A deep network solution for attention and aesthetics aware photo cropping," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1531–1544, 2018.

[3] W. Wang, X. Lu, J. Shen, D. Crandall, and L. Shao, "Zero-shot video object segmentation via attentive graph neural networks," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, pp. 9235–9244.

[4] X. Dong, J. Shen, D. Wu, K. Guo, X. Jin, and F. Porikli, "Quadruplet network with one-shot learning for fast visual object tracking," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3516–3527, 2019.

[5] Z. Liang and J. Shen, "Local semantic siamese networks for fast tracking," *IEEE Transactions on Image Processing*, vol. 29, pp. 3351–3364, 2019.

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

[7] K. Yang, X. Hu, L. M. Bergasa, E. Romera, X. Huang, D. Sun, and K. Wang, "Can we pass beyond the field of view? panoramic annular semantic segmentation for real-world surrounding perception," in *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2019, pp. 446–453.

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

[9] 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)*. IEEE, 2017, pp. 2261–2269.

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

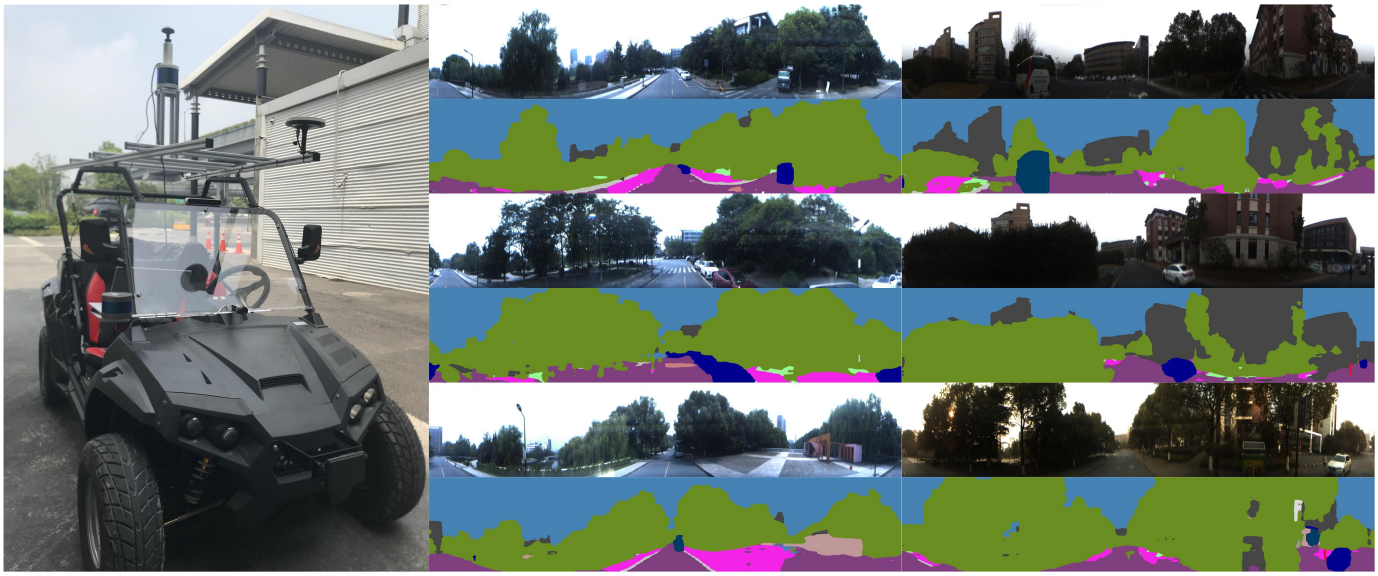
[11] G. Neuhold, T. Ollmann, S. R. Bulò, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 5000–5009.

[12] 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.

[13] 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)*. IEEE, 2019, pp. 12 599–12 608.

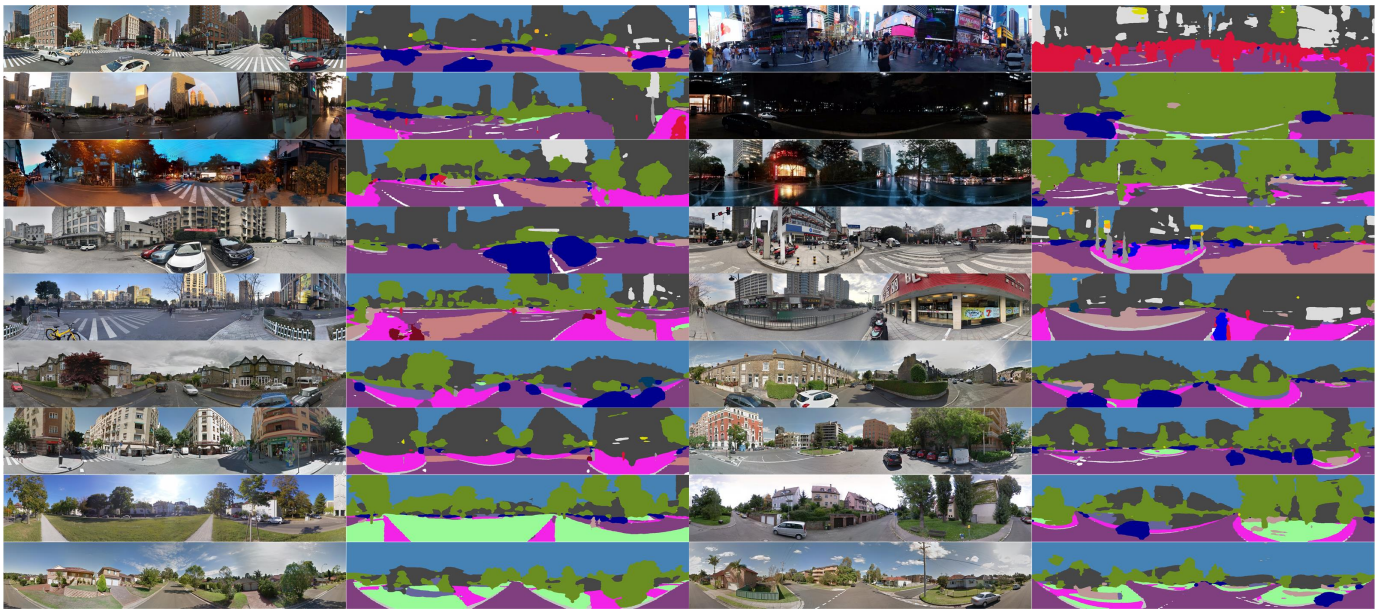
[14] K. Yang, X. Hu, H. Chen, K. Xiang, K. Wang, and R. Stiefelhagen, "Ds-pass: Detail-sensitive panoramic annular semantic segmentation through swaftnet for surrounding sensing," in *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2020.

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



(a) Instrumented vehicle (b) Summer panorama (c) Winter panorama

Fig. 12. Qualitative examples of omnisupervised omnidirectional semantic segmentation in cross-season real-world scenes: (b) summer and (c) winter results where the system is deployed on (a) the instrumented vehicle with a panoramic annular lens.



(a) Input (b) Prediction (c) Input (d) Prediction

Fig. 13. Qualitative examples of omnisupervised omnidirectional semantic segmentation by virtually navigating the instrumented vehicle around the world: (a)(c) Inputs, (b)(d) Predictions. From top to bottom: New York, Beijing, Shanghai, Changsha, Hangzhou, Huddersfield, Madrid, Karlsruhe and Sydney.

[16] K. Yang, X. Hu, L. M. Bergasa, E. Romera, and K. Wang, "Pass: Panoramic annular semantic segmentation," *IEEE Transactions on Intelligent Transportation Systems*, 2019.

[17] S. Yogamani, C. Hughes, J. Horgan, G. Sistu, S. Chennupati, M. Uricar, S. Milz, M. Simon, K. Amende, C. Witt *et al.*, "Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, pp. 9307–9317.

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

[19] 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 *2019 British Machine Vision Conference (BMVC)*, 2019, p. 31.

[20] L. Deng, M. Yang, H. Li, T. Li, B. Hu, and C. Wang, "Restricted deformable convolution-based road scene semantic segmentation using surround view cameras," *IEEE Transactions on Intelligent Transportation Systems*, 2019.

[21] 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)*. IEEE, 2017, pp. 231–236.

[22] 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)*. IEEE, pp. 3532–3540.

[23] Y. Xu, K. Wang, K. Yang, D. Sun, and J. Fu, "Semantic segmentation of panoramic images using a synthetic dataset," in *Artificial Intelligence and Machine Learning in Defense Applications*, vol. 11169. International Society for Optics and Photonics, 2019, p. 111690B.

[24] Y. Ye, K. Yang, K. Xiang, J. Wang, and K. Wang, "Universal semantic

segmentation for fisheye urban driving images," in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2020.

[25] E. Romera, L. M. Bergasa, K. Yang, J. M. Alvarez, and R. Barea, "Bridging the day and night domain gap for semantic segmentation," in *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2019, pp. 1312–1318.

[26] H. Porav, T. Bruls, and P. Newman, "Don't worry about the weather: Unsupervised condition-dependent domain adaptation," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 33–40.

[27] L. Sun, K. Wang, K. Yang, and K. Xiang, "See clearer at night: towards robust nighttime semantic segmentation through day-night image conversion," in *Artificial Intelligence and Machine Learning in Defense Applications*, vol. 11169. International Society for Optics and Photonics, 2019, p. 111690A.

[28] 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 (CVPR)*. IEEE, 2018, pp. 4119–4128.

[29] 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.

[30] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *2016 International Conference on Learning Representations (ICLR)*, 2016.

[31] 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)*. IEEE, 2017, pp. 6230–6239.

[32] 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.

[33] 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 (CVPR)*. IEEE, 2018, pp. 3684–3692.

[34] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019, pp. 3141–3149.

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

[36] 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)*. IEEE, 2019, pp. 1440–1444.

[37] Y. Pang, Y. Li, J. Shen, and L. Shao, "Towards bridging semantic gap to improve semantic segmentation," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, pp. 4229–4238.

[38] 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.

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

[40] M. Treml, J. Arjona-Medina, T. Unterthiner, R. Durgesh, F. Friedmann, P. Schuberth, A. Mayr, M. Heusel, M. Hofmarcher, M. Widrich *et al.*, "Speeding up semantic segmentation for autonomous driving," in *ML-ITS, NIPS Workshop*, vol. 2, 2016, p. 7.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

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

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

[47] B. Pan, J. Sun, A. Andonian, A. Oliva, and B. Zhou, "Cross-view semantic segmentation for sensing surroundings," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4867–4873, 2020.

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

[49] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognition Letters*, vol. 30, no. 2, pp. 88–97, 2009.

[50] L. Wang and K.-J. Yoon, "Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks," *arXiv preprint arXiv:2004.05937*, 2020.

[51] C. Gong, X. Chang, M. Fang, and J. Yang, "Teaching semi-supervised classifier via generalized distillation," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2018, pp. 2156–2162.

[52] J. Vongkulbhisal, P. Vinayavekhin, and M. Visentini-Scarzanella, "Unifying heterogeneous classifiers with distillation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019, pp. 3170–3179.

[53] L. Xiang and G. Ding, "Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification," *arXiv preprint arXiv:2001.01536*, 2020.

[54] Y. Liu, X. Dong, W. Wang, and J. Shen, "Teacher-students knowledge distillation for siamese trackers," *arXiv preprint arXiv:1907.10586*, 2019.

[55] A. Wu, W.-S. Zheng, X. Guo, and J.-H. Lai, "Distilled person re-identification: Towards a more scalable system," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019, pp. 1187–1196.

[56] P. Meletis and G. Dubbelman, "Training of convolutional networks on multiple heterogeneous datasets for street scene semantic segmentation," in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 1045–1050.

[57] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019, pp. 3962–3971.

[58] J. Xie, B. Shuai, J.-F. Hu, J. Lin, and W.-S. Zheng, "Improving fast segmentation with teacher-student learning," in *2018 British Machine Vision Conference (BMVC)*, 2018, p. 205.

[59] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured knowledge distillation for semantic segmentation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019, pp. 2599–2608.

[60] T. He, C. Shen, Z. Tian, D. Gong, C. Sun, and Y. Yan, "Knowledge adaptation for efficient semantic segmentation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019, pp. 578–587.

[61] Y. Zhang, P. David, and B. Gong, "Curriculum domain adaptation for semantic segmentation of urban scenes," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 2039–2049.

[62] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *Proceedings of the International Conference on Machine Learning (ICML)*, vol. 80, 2018, pp. 1994–2003.

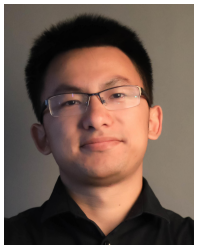
[63] Y. Zou, Z. Yu, X. Liu, B. V. Kumar, and J. Wang, "Confidence regularized self-training," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, pp. 5981–5990.

[64] L. Sun, K. Yang, X. Hu, W. Hu, and K. Wang, "Real-time fusion network for rgb-d semantic segmentation incorporating unexpected obstacle detection for road-driving images," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5558–5565, 2020.

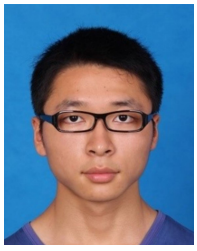
[65] T. Kalluri, G. Varma, M. Chandraker, and C. Jawahar, "Universal semi-supervised semantic segmentation," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, pp. 5258–5269.

[66] 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.

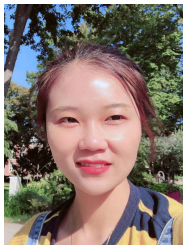
[67] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi, "Visual place recognition with repetitive structures," in *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2013, pp. 883–890.
[68] M. Zhu, W. Wang, B. Liu, and J. Huang, "A fast image stitching algorithm via multiple-constraint corner matching," *Mathematical Problems in Engineering*, vol. 2013, 2013.
[69] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.
[70] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *2015 International Conference on Learning Representations (ICLR)*, 2015.
[71] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 5122–5130.



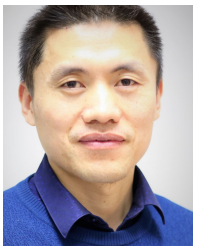
Kailun Yang received the B.S. degree in Measurement Technology and Instrument from Beijing Institute of Technology (BIT) and the second degree in Economics from Peking University (PKU) in June 2014. He received the Ph.D. degree in Information Sensing and Instrumentation from the State Key Laboratory of Modern Optical Instrumentation, Zhejiang Univeristy (ZJU) in June 2019. He has performed a Ph.D. internship at the RobeSafe Lab, University of Alcalá (UAH) from September 2017 to September 2018. He is currently a postdoctoral researcher at the Computer Vision for Human-Computer Interaction Lab at the Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology (KIT). His research interests include real-time computer vision, optical sensing and their applications for intelligent vehicles and real-world navigation assistance systems for the visually impaired. For more information, visit his Website: http://www.yangkailun.com/.



Xinxin Hu received the B.S. degree from the College of Optical Science and Engineer, Zhejiang University in 2017. He is studying for a master's degree in Instrument and Meter Engineering, at the State Key Laboratory of Modern Optical Instrumentation, Zhejiang University. He has internship experience in central media research institute of Huawei and ArcSoft, mainly engaged in knowledge distillation and portrait segmentation. His research interests include optical detection, 3D vision, semantic segmentation, knowledge distillation and indoor navigation.



Yicheng Fang received the B.S. degree from Yanshan University, China in 2018. She performed a research on tapered fiber optic temperature sensor during the undergraduate period. Currently, she is a master student at the State Key Laboratory of Modern Optical Instrumentation, Zhejiang University, China. Her present research interest and knowledge are mainly centered around visual place recognition for visually impaired people, mobile robots and intelligent vehicles.



Kaiwei Wang is currently a full professor at the State Key Laboratory of Modern Optical Instrumentation, and the Deputy Director of the National Optical Instrument Engineering Research Center at Zhejiang University. He received a B.S. degree in 2001 and a Ph.D. degree in 2005 respectively, both from Tsinghua University. In October 2005, he started his postdoctoral research at the Center of Precision Technologies (CPT) of Huddersfield University, funded by the Royal Society International Visiting Postdoctoral Fellowship and the British Engineering Physics Council. He joined Zhejiang University in February 2009 and has been mainly researching on intelligent optical sensing technology and visual assisting technology for the visually impaired. Up to date, he owns 60 patents and has published more than 150 refereed research papers. For more information, visit his Website: http://wangkaiwei.org/.

Rainer Stiefelhagen received his Diplom (Dipl.-Inform) and Doctoral degree (Dr.-Ing.) from the Universität Karlsruhe (TH) in 1996 and 2002, respectively. He is currently a full professor for "Information technology systems for visually impaired students" at the Karlsruhe Institute of Technology (KIT), where he directs the Computer Vision for Human-Computer Interaction Lab at the Institute for Anthropomatics and Robotics as well as KIT's Study Center for Visually Impaired Students. His research interests include computer vision methods for visual perception of humans and their activities, in order to facilitate perceptive multimodal interfaces, humanoid robots, smart environments, multimedia analysis and assistive technology for persons with visual impairments.