In Defense of Multi-Source Omni-Supervised Efficient ConvNet for Robust Semantic Segmentation in Heterogeneous Unseen Domains

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

***Abstract*— Semantic segmentation renders a unified way of surrounding perception, where most of driving scene detection tasks can be covered by running a single efficient ConvNet through a forward pass. However, current frameworks posit the closed-world paradigm expressed as a single source of distribution over a predetermined set of visual classes, forgetting that a deep model must be deployed in the wild facing unseen domains and unforeseen hazards. In spite of being accurate in its comfort zone, the segmentation model may not generalize well to a new domain. In addition, a model trained with single dataset is heavily limited in terms of recognizable classes. In this paper, we propose an omni-supervised learning framework for semantic segmentation which is able to leverage heterogeneous data sources. Our omni-supervised training framework incorporates all available labeled and unlabeled data, meanwhile bridges multiple training sets to be capable of recognizing more classes that are needed for autonomous navigation application at hand in the new domain. A comprehensive variety of experiments shows that with the proposed multi-source omni-supervised learning solution, an efficient ConvNet like our ERF-PSPNet attains significant robustness gains in open domains that are of critical relevance to real deployment of vision algorithms. Our approach surpasses the state of the art on the highly unconstrained PASS and IDD20K datasets.**

I. INTRODUCTION

Semantic segmentation supposes a unified manner of surrounding scene sensing to cover most of perception needs of Intelligent Vehicles (IV) [1][2]. Emergence of large natural datasets and architectural advances of deep models have reinforced the excellence of Convolutional Networks (ConvNets) at this task, allowing them to predict pixel-wise semantics over a predetermined set of visual classes both accurately and efficiently with a single forward pass [3].

Unfortunately, existing frameworks assume the closed-world evaluation with a single-source setting, posing overwhelming difficulties for real-life applications, as a deep model may be biased towards a comfortable domain while not generalizing in the wild, yet a single training dataset is far from being comprehensive [4]. When a trained model is taken from its comfort zone to an unseen domain, the performance usually declines dramatically and even catastrophically due to the large imagery gap between real-world heterogeneous domains [5][6]. Besides, a model learned with a single dataset is heavily limited in terms of recognizable classes, which are only a subset of semantics required to gain a complete scene understanding. The comprehension of a vehicle's surroundings becomes even more challenging in specific situations such as intersections, roundabouts or unconstrained environments with exceedingly diverse dynamic traffic participants whose behaviors are highly unpredictable [3]. To promote the detection of unforeseen hazards, one has to perform post-processing or run another model trained on additional sets, which both significantly increase inference latency and computation complexity [7].

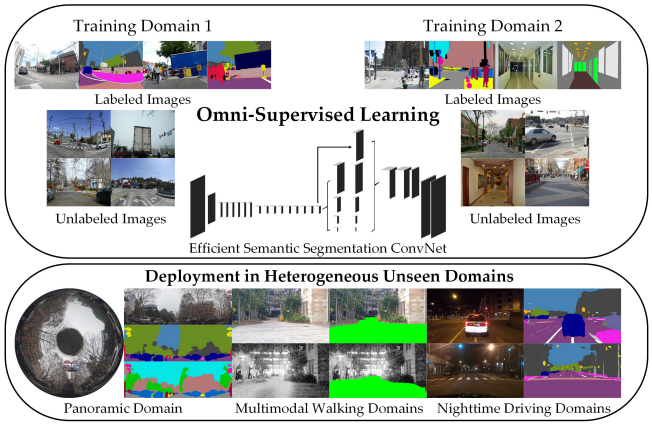


Fig. 1. Overview of the proposed omni-supervised solution: an efficient ConvNet is trained using both labeled and unlabeled images, yielding a single unified model that is robust across heterogeneous unseen domains.

In the vision community, there are already a large body of semantic segmentation datasets [8][9][10] available to produce trained models, generally practitioners only use one due to the classes contradictions among the datasets. For example, road surfaces are simply labeled as road and sidewalk in Cityscapes [4], but in Mapillary Vistas [8], they are labeled as road, sidewalk with curbs between them, and additional roadway classes like crosswalks. In addition, riders in Cityscapes would be divided into motorcyclists and bicyclists if from Vistas. There are novel classes like auto-rickshaws that are not existing in Cityscapes (European streets) but widespread in the unstructured IDD dataset (Asia) [10]. While one can leverage model variants trained using multiple datasets or domain bridges [5], this is unsatisfactory for IV applications as the same model should be directly deployable in a broad spectrum of environments (see Fig. 1).

There are proposals that aim to solve the conflictions between the class hierarchies [11][12], but only the annotated images from these large-scale databases are fed with the model to train in a fully-supervised way, leaving the diversity

This work has been partially funded by the Federal Ministry of Labour and Social Affairs from the compensation fund. This work has also been supported by Hangzhou SurImage Technology Co., Ltd. and Hangzhou KrVision Technology Co., Ltd. (krvision.cn).

[1]K. Yang and R. Stiefelhagen are with Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Germany {kailun.yang, rainer.stiefelhagen}@kit.edu

[2]X. Hu and K. Wang are with National Optical Instrumentation Engineering Technology Research Center, Zhejiang University, China {hxx_zju, wangkaiwei}@zju.edu.cn

implicated in the unlabeled data unexploited. On the other hand, since efficient ConvNets are usually benchmarked against datasets with a limited set of classes, it was stated that they cannot resolve the real-world complexity burdened by the increased number of visual classes on the semantically-connected datasets [12].

To achieve real-world robustness in efficient architectures, this paper proposes an omni-supervised learning framework for semantic segmentation which exploits multiple heterogeneous datasets. Following the concept of data distillation for omni-supervised learning [13], the learner in our framework incorporates all available labeled and unlabeled images. Meanwhile, the proposed framework bridges multiple domains in a general way without having to regulate the complex class hierarchies among datasets. Vitally, our experiments show that an omni-supervised efficient ConvNet delivers both rich sets of recognizable classes and high generalization capacities empowered by multi-source data and data distillation. To the best of our knowledge, this is the first time that omni-supervised learning has been leveraged to address navigational scene parsing.

More precisely, we involve large-scale Mapillary Vistas [8] and ADE20K/IDD20K [9][10] datasets to perform learning and validation. Taking a key step further, we verify the performances in the totally unseen domains of Panoramic Annular Semantic Segmentation (PASS) [2][14] and Gardens Point [15]. For Gardens Point set, we create pixel-wise semantic traversable area annotations to facilitate evaluation. With a comprehensive variety of experiments on these navigational scene parsing datasets, the ultimate goal is to yield a single model that works robustly across different domains, including known domains and more crucially new, open domains that are of critical relevance to real deployment of efficient ConvNets like ERF-PSPNet [1][3].

We also investigate the generalization benefits in diverse road-sensing datasets [10][16][17], showing that our approach supports robust semantic perception even in adverse conditions such as the night scenarios in multimodal walking and nighttime driving domains (see Fig. 1). Our proposal surpasses the state of the art on IDD20K and PASS datasets, both of which reflect the challenges of IV applications in unconstrained street scenes. Our codes and datasets will be made publicly available at [1].

II. RELATED WORK

A. Architectural Advances of Semantic Segmentation

Fully Convolutional Network (FCN) [18] opens the vista of deep end-to-end semantic segmentation, whose performance was exceeded by subsequently appeared DRNet (with dilated convolution) [19] and PSPNet (with pyramid pooling) [20]. In [7], dilated convolutions were extended to hierarchical architectures to predict pixel-wise specular semantics such as water hazards for wearable robotics. In road scene understanding-desired IV applications, ERFNet [21], ERF-PSPNet [1][3], SwiftNet [22] were developed to boost the real-time performance. Attention connections (SwaftNet) [2] were put forward to improve the detail-sensitivity, rendering surroundings segmentation both swiftly and accurately, without sitting on one side of the balance.

B. Domain Adaptation and Semi-Supervised Learning

To address the dearth of large-scale database that is critical to produce robust segmentation models, synthetic data have been frequently used to augment the training set [3][23]. The mismatch between synthetic and real scenes arouses an army of domain adaptation researches [5][6][16][24]. A critical subset of these proposals particularly aims to rectify the imagery gap and improve the performance in adverse conditions (e.g. nighttime) based on curriculum learning [16] or unsupervised image translation [5][6], assuming clear boundaries between different domains, such as the day and night. Another appealing line is semi-supervised learning that mitigates the deficiency of pixel-aware labels by using weak supervision like bounding-box annotations [25].

While previous adaptation approaches pride their performance in discrete domains [3][5][6], we aim to produce a unified model that generalizes in open domains previously unseen during the training stage without defining any boundaries. Sharing the similar spirit for deployment regardless of the training domain, [26][27] explored aggressive data augmentation and universal semi-supervised knowledge aggregation but only verified the generalizability in scenarios with limited heterogeneity. In this work, following the concept of omni-supervised learning [13], we extend to multi-source semantic segmentation by leveraging full pixel-wise labeled and unlabeled data from heterogeneous training sets for efficient ConvNet deployable in countless autonomous transportation applications. Overall, while we have witnessed significant progress of unsupervised domain adaptation and semi-supervised learning in this field [28][29][30], it is expected that one model can only be trained once yet it could generalize well in new coming scenarios. Thereby, our approach is closer to domain generalization, whose key difference to domain adaptation is that no training samples are available in the target domain.

C. Semantic Segmentation with Diverse Supervision

Learning semantic segmentation with diverse supervision has been previously addressed in [14][23]. In [14], style-transferred, distorted and high-definition images are blended in training to robustify against blurs and distortions. In [23], synthetic fish-eye forward-view images and real-world surround-view images are incorporated by tuning the hybrid loss functions, which does not fit for learning with multiple real-world sets as it requires storing private domain-specific statistics, but our aim is to yield a single model that supports reliable deployment across domains.

In contrast, due to the proliferation of large real-world finely-annotated segmentation databases, the relations between different semantic classes hierarchies have been carefully entangled in [11][12], to facilitate end-to-end super-

[1]Datasets and codes of omni-supervised efficient ConvNet: https://github.com/elnino9ykl/OmniSupervised-ConvNet

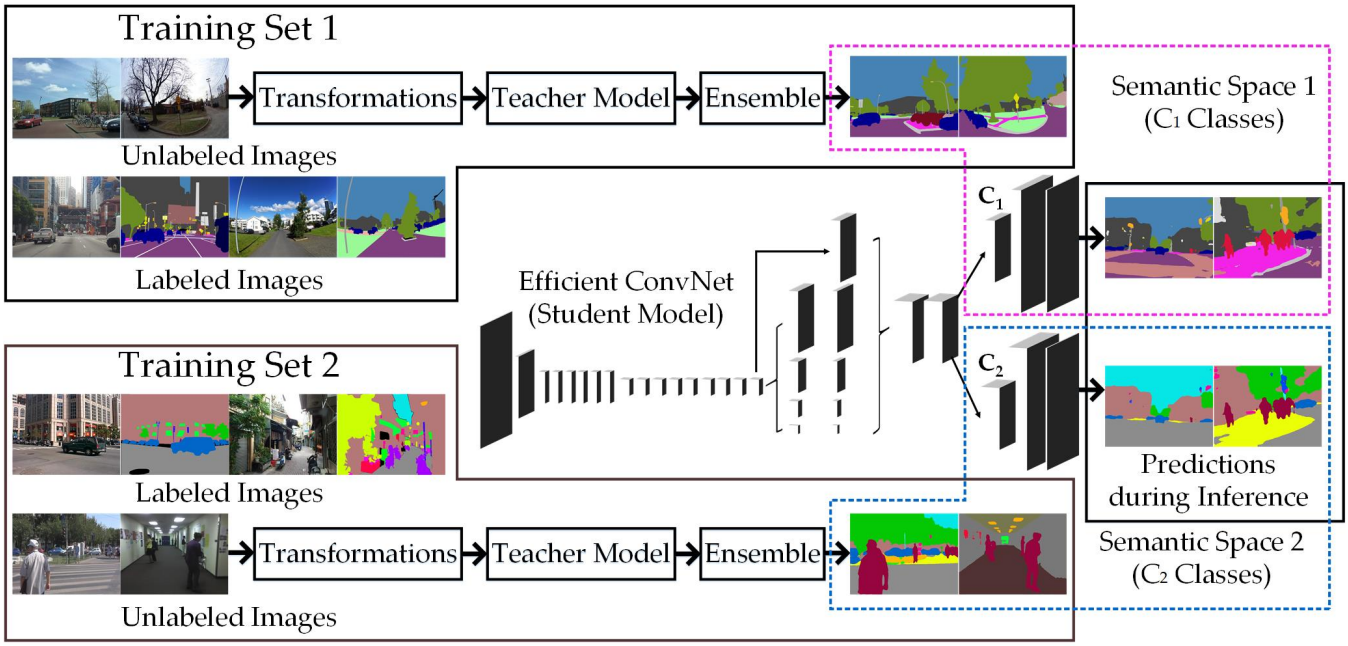


Fig. 2. The proposed omni-supervised learning framework. During training, both labeled and unlabeled images from heterogeneous datasets are exploited, where annotations for unlabeled images are automatically generated by creating ensembles of a teacher model's predictions on multiple transformations of the input. During deployment, the yielded student model is not only efficient and robust, but also capable of recognizing diverse sets of semantic classes.

vised training on multiple datasets. In their experiment, it was highlighted that efficiency-oriented networks are not able to deal with the complex class hierarchy of the joint dataset. For this reason, they stated that contemporary efficient ConvNets are limited in terms of handling a large number of heterogeneous semantic classes [12], which are required to fully understand real-world unconstrained surroundings. However, our experiments suggest that with the proposed framework, an efficient segmentation ConvNet is not only capable of recognizing richer sets of classes, but also empowered with higher robustness in heterogeneous domains.

III. FRAMEWORK

A. Overview

The diagram of the proposed omni-supervised solution is depicted in Fig. 2. During the training phase, we create ensembles for a teacher model's predictions on the transformations of the unlabeled images. We exploit both manually labeled and the extra generated annotations to train the efficient student network. During deployment, the yielded robust ConvNet can run in real time on autonomous navigation systems, while delivering diverse sets of recognizable classes required to fully understand real-world unconstrained surroundings. In the following sub-sections, we will describe in detail the training (Section III-B) and deployment (Section III-C) of efficient ConvNet for robust semantic segmentation across heterogeneous domains.

B. Training Stage

As there are countless semantic segmentation datasets available in the vision community, we exploit N large-scale datasets for training, each of which D_i $(i = 1{\sim}N)$ corresponds to a specific domain, having labeled samples S_{il} and unlabeled samples S_{iu}. The annotations for the labeled samples are A_{il}, with a semantic class space C_i. To train an efficient semantic segmentation ConvNet F, the conventional strategy is to learn the mapping represented by the following equation:

$$F\left(S_{il}\right) \Longrightarrow A_{il}(C_i)$$

The efficient segmentation model F, usually separated as encoder-decoder in a sequential way, can be re-separated into a feature model F_e that first predicts high-level abstract features and a pixel-wise classification model F_c that maps the feature map to the specific semantic space, formally:

$$F\left(S_{il}\right) = F_c\left[F_e\left(S_{il}\right)\right] \Longrightarrow A_{il}(C_i)$$

The purpose of our work is to train a single unified model that is useful across different domains, but the semantic spaces in different datasets are incompatible. For ease of notation, in the case of two domains, $C_1 \neq C_2$, which means that the classes are heterogeneous and classes numbers are usually not equivalent across domains, although they are partially overlapping with each other. The labels definitions often encode relationships that positively reinforce the robustness and generalizability of feature representations when learning across different domains. In this sense, it is fruitful to learn with multiple datasets, even their semantic spaces are with high heterogeneity. Considering the two domains case, the training target can be modified into:

$$F_{c1}\left[F_e\left(S_{1l}\right)\right], F_{c2}\left[F_e\left(S_{2l}\right)\right] \Longrightarrow A_{1l}(C_1), A_{2l}(C_2),$$

where a single efficient model is armed with two classification sub-models, which only slightly increases the computational overhead, but largely enriches recognizable classes to fully understand real-world surroundings, and the two domains case can be easily scaled up to multiple datasets.

We further extend the solution to omni-supervised setting by incorporating unlabeled data for training. Although top-performance segmentation models are not efficient and lightweight enough to be deployed in autonomous transportation systems, their produced semantic maps are highly qualified and finely grained, and an ensemble of the outputs of a single model run on multiple transformed copies of an unlabeled example is known to improve the performance [20], leading to very accurate results that can be trusted for distillation. Unlike model distillation that entails re-training different heavy networks, in this research we follow the concept of data distillation [13] by using a state-of-the-art architecture that has been independently trained in each domain, which is more flexible in the large databases case. The yielded domain-specific teacher model F_{ti} is responsible for automatically generating annotations for the unlabeled samples:

$$A_{iu}(C_i) = \biguplus_{j=1}^{M} \left[F_{ti}\left(T_j\left(S_{iu} \right) \right) \right],$$

where A_{iu} are generated annotations for unlabeled samples, which also fall in the semantic space C_i, $T_j(j = 1{\sim}M)$ denotes a transformation of an input image, and $\biguplus$ represents the created ensemble for the predictions of the teacher model through averaging, weighting or probability maps aggregating. For semantic segmentation, diverse geometric and textural transformations can be used, such as horizontal flipping, resizing and color jittering.

Finally, the newly generated annotations of unlabeled sources are blended with the manually annotated set to train the efficient ConvNet:

$$F_{ci}\left[F_e\left(S_{il}, S_{iu} \right) \right] \Longrightarrow \left(A_{il}, A_{iu} \right)(C_i)$$

C. Deployment Stage

After training, the ConvNet is ready for being applied in previously unseen domains, while no transformation is needed in the deployment phase. The resulted learner is a single model, which maintains the efficiency and simplicity as in common case of semantic perception systems, but it possesses several important benefits. First, since it has been exposed to diverse scenes from multiple datasets, the generalizability and inherent robustness of the model have been significantly enhanced. Second, the model is able to deliver more detectable semantics:

$$\sum_{i=1}^{N} F_{ci}\left[F_e\left(S_n \right) \right] = \sum_{i=1}^{N} \left(M_i(C_i) \right),$$

where for a new sample S_n, N semantic maps will be generated, each of which M_i corresponds to a semantic space C_i, supposing a very rich resource of mutually complementary information for upper-level navigational applications.

IV. EXPERIMENTS AND RESULTS

A. Training Datasets

We perform experiments in the case of learning with two datasets, but it can be easily scaled up to more image sources. We employ the challenging Mapillary Vistas [8] and ADE20K [9], two of the largest and richest scene parsing datasets in the computer vision community today. We use the most frequent and navigation-related categories for training, namely 25 classes of Vistas and 50 classes of ADE20K to prevent model underfitting and facilitate fair evaluation across domains. Vistas exhibits an appealing diversity with images captured in multiple continents and various viewpoints, such as from road (vehicles), sidewalk (pedestrians) and off-road views. It splits into 18000/2000/5000 images in the train, validation and test subsets. ADE20K comprises images from both indoor and outdoor, splitting into 20210/2000/3352 samples for training, validation and testing. This combination has implications for a wide variety of transportation applications, such as mobility assistance for the visually impaired [1]. Since the annotations for the testing images are not publicly available, we evaluate on the validation sets with equal amount (2000 images).

While these two datasets provide a rich ontology with many factors of variations, we also leverage the recently updated IDD20K dataset [10] and consider the combination of training with Vistas and IDD20K. IDD20K has 14027/2036/4038 images for training/validation/testing on 26 classes, which covers extremely unstructured environments. This combination is more oriented to autonomous driving, allowing us to study the generalization benefits of our approach for unconstrained surroundings that raise enormous challenges to the robustness of semantic segmentation.

B. Training Setups

We experiment with our real-time ConvNet ERF-PSPNet [1][2], which was developed in previous work for navigation assistance, but the study is conducted in a general way that is applicable to any deep efficient architecture. The models are trained under Adam optimization with a weight decay of $2{\times}10^{-4}$ and a starting learning rate of $5{\times}10^{-4}$ that decreases exponentially over 200 epochs. Samples are fed with a batch size of 24 and a resolution of 512×512 as a balance between the two heterogeneous training sets (Vistas+ADE20K). We only use random horizontal flipping to transform the inputs, where other data augmentation and domain adaptation strategies [14] that have been proven beneficial to generalization capacity are kept out. This is more realistic as under common cases, the domain knowledge about the deployment environment is unaccessible. We employ the standard mean Intersection-over-Union (mIoU) as the evaluation metric.

We use the state-of-the-art PSPNet50 [20] as the teacher model, which has 67.1% mIoU on Vistas and 48.7% on ADE20K validation sets. Our omni-supervised learning setting is based on creating annotations for the 5000 and 3352 unlabeled images from the test subsets. We use horizontal

TABLE I
COMPARISON OF SEMANTIC SEGMENTATION ACCURACY MEASURED IN MIOU ACROSS HETEROGENEOUS DOMAINS.
TRAINING RESOLUTION: 512×512.

Network	Trained Domains (Validation Sets)		Unseen Domains (for Deployment)	
	Mapillary Vistas	ADE20K	PASS	Gardens Point
ERF-PSPNet (Vistas-trained)	58.9%	NA	27.4%	68.4%
ERF-PSPNet (ADE20K-trained)	NA	38.8%	17.7%	71.7%
ERF-PSPNet (jointly-trained)	54.1%	44.2%	32.6%	76.6%
Omni-Supervised ERF-PSPNet	53.4%	44.1%	**37.8%**	**80.0%**

mirroring, 2× and 0.5× scaling for each unlabeled image, resulting in 4 transformed duplicates including the original one. We ensemble the teacher model's predictions by aggregating the probability maps for these copies to form as the annotation of the unlabeled set for data distillation. When training on multiple datasets, each iteration contains a forward pass and a backward pass per dataset using cross-entropy loss functions.

C. Baselines

The training results are shown in Table I, where our ERF-PSPNet achieves 58.9% and 38.8% on the validation sets when trained on Mapillary Vistas and ADE20K independently. For the IDD20K dataset and the Vistas+IDD20K cases, we train on 1024×512 as they both support high-definition inputs while other training implementation details are the same as the Vistas+ADE20K combination. Our ERF-PSPNet achieves 63.2%/64.2% on the IDD20K validation dataset without/with Vistas-supervision, both surpassing the results of state-of-the-art efficient networks such as DRNet [19] and ERFNet [21], as displayed in Table II. Compared with the Universal Semi-supervised Semantic Segmentation (USSS) approach [27] that incorporates part of images from both IDD and Cityscapes for training, our score is also higher as we are able to leverage the full IDD20K-based supervision. However, in spite of the decent mIoU numbers on the validation sets, there is a large accuracy downgrade when a single-source trained model is taken to previously unseen domains such as the panoramic domain and the multimodal walking domain as it can be seen in Table I, Table III and Table IV. The main aim of our experiments is to answer whether the proposal benefits the generalizability. The short answer is yes as shown in the Tables. In the following subsections, we will examine in detail the generalization gains in various unseen domains.

TABLE II
ACCURACY ANALYSIS ON IDD20K DATASET [10].
TRAINING RESOLUTION: 1024×512.

Network	IDD20K	Vistas
DRNet (ResNet18) [19]	52.2%	NA
USSS (ResNet18) [27]	27.5%	NA
USSS (ResNet50) [27]	55.1%	NA
ERFNet [21]	55.4%	NA
Our ERF-PSPNet (Vistas-trained)	NA	61.6%
Our ERF-PSPNet (IDD-trained)	63.2%	NA
Our ERF-PSPNet (Jointly-trained)	**64.2%**	**63.0%**

D. Panoramic Domain

For panoramic images, semantic segmentation is required to cover the field of view as wide as 360°, which has important implications as comprehensive perception of the entire surrounding is necessary for IV applications. In this work, we use the Panoramic Annular Semantic Segmentation (PASS) dataset [2][14] which comprises 400 images for testing with pixel-accurate annotations on 6 navigation-critical classes: Car, Road, Sidewalk, Crosswalk, Curb and Person.

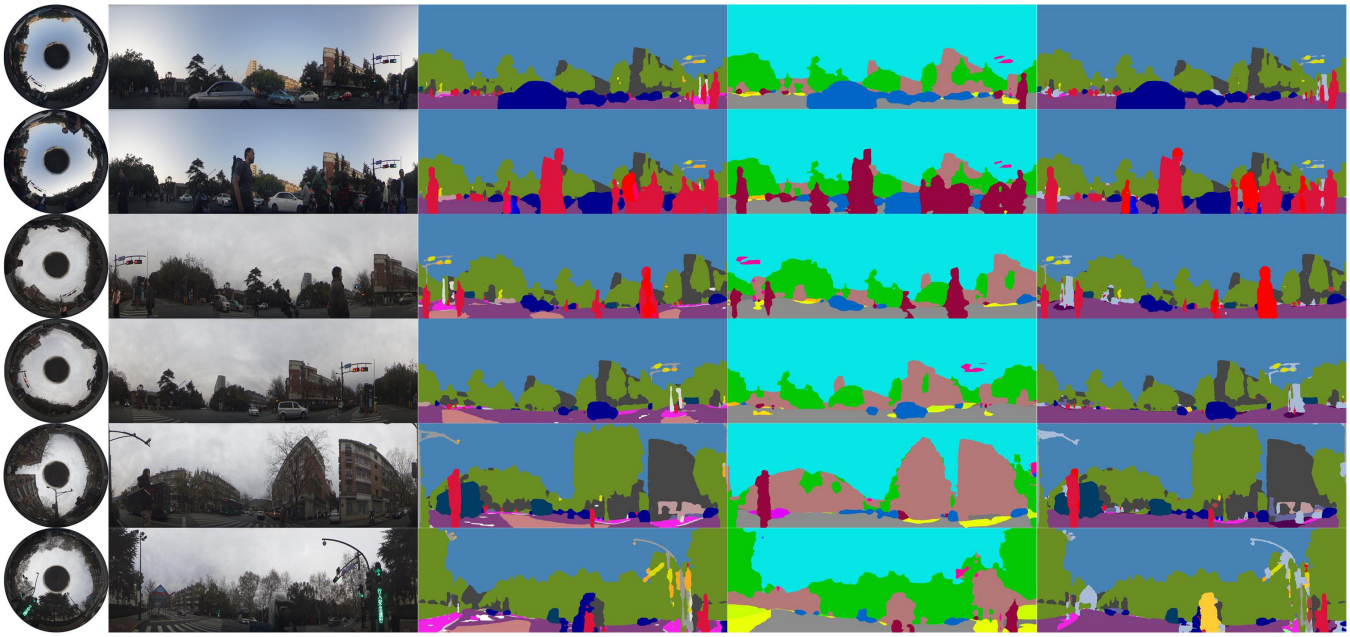
As shown in Table I and Table III, the ERF-PSPNet trained on a single dataset performs poorly in the panoramic domain as the mIoU numbers (27.4% and 17.7%) are below 30.0%. When performing training jointly with Vistas and ADE20K, the accuracy increases to 32.6%, demonstrating the benefit of diverse supervision from the two heterogeneous datasets. This is because when trained with heterogeneous datasets, the network learns to focus on relevant features for both tasks and thus gains robustness. The training decreases the accuracy on Vistas compared to the single-source Vistas-training, which proves that the joint-training highly prevents overfitting, as the accuracies in multiple unseen domains significantly improve. The joint-training also increases the accuracy on ADE20K because it contains many street-scene images that benefit from Vistas-based supervision.

The omni-supervised learning further improves the accuracy, reaching 37.8% on PASS, verifying the benefit of variety implicated in the unlabeled data. This score surpasses most of the previous attempts on this dataset including ERFNet, PSPNet18, ERF-APSPNet [14] and SwiftNet [22]. While these networks are trained with aggressive data augmentation and style transfer-based domain adaptation strategies [2], our approach is more realistic as generally the style of the target domain is unknown. This result also demonstrates that omni-supervised learning is beneficial for the challenging omni-directional semantic segmentation task.

PASS dataset is a highly unconstrained domain as in panoramic imagery, traffic participants with diverse orientations can be simultaneously observed (see examples in Fig. 3). In addition, sometimes there are many close participants present in the image from PASS dataset. For these reasons, our solution combining Vistas and IDD20K for training significantly improves the performance as IDD20K embraces highly unstructured roads while Vistas offers the high diversity. While IDD20K-trained and Vistas-trained models yield 20.1% and 27.4% on PASS respectively, the joint-training boosts the accuracy to 41.0%, outperforming all previous networks attempted on this dataset [2]. This

TABLE III
ACCURACY ANALYSIS ON PANORAMIC ANNULAR SEMANTIC SEGMENTATION (PASS) DATASET [2][14].
ALL NETWORKS ARE TESTED BY VIEWING THE PANORAMA AS A SINGLE SEGMENT EXCEPT THOSE WITH CROSS-SEGMENT PADDING.

Network	Car	Road	Sidewalk	Crosswalk	Curb	Person	mIoU
ERFNet [21]	70.0%	57.3%	25.4%	22.9%	15.8%	15.3%	34.3%
PSPNet (ResNet18) [20]	64.1%	67.7%	31.2%	15.1%	17.5%	12.8%	34.8%
ERF-APSPNet [14]	72.3%	71.4%	32.6%	5.6%	16.3%	14.5%	35.5%
SwiftNet [22]	67.5%	70.0%	30.0%	21.4%	21.9%	13.7%	37.4%
SwaftNet [2]	76.4%	64.1%	33.8%	9.6%	26.9%	18.5%	38.2%
ERF-PSPNet (Vistas)	57.2%	55.2%	17.9%	13.5%	11.7%	8.8%	27.4%
ERF-PSPNet (ADE20K)	36.9%	50.7%	14.9%	0.0%	0.0%	3.5%	17.7%
ERF-PSPNet (ADE20K+Vistas)	64.8%	68.7%	28.6%	4.6%	17.3%	11.5%	32.6%
Omni-Supervised	**68.4%**	**74.0%**	**39.6%**	**11.6%**	**18.7%**	**14.8%**	**37.8%**
Omni-Supervised (with cross-segment padding)	**88.0%**	**79.5%**	**41.7%**	**57.9%**	**32.7%**	**52.4%**	**58.7%**
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 (IDD20K+Vistas, with cross-segment padding)	**91.0%**	**82.5%**	**56.8%**	**56.9%**	**38.2%**	**74.1%**	**66.6%**



(a) Raw (b) Unfolded panorama (c) Prediction in Vistas space (d) Prediction in ADE20K space (e) Prediction in IDD20K space

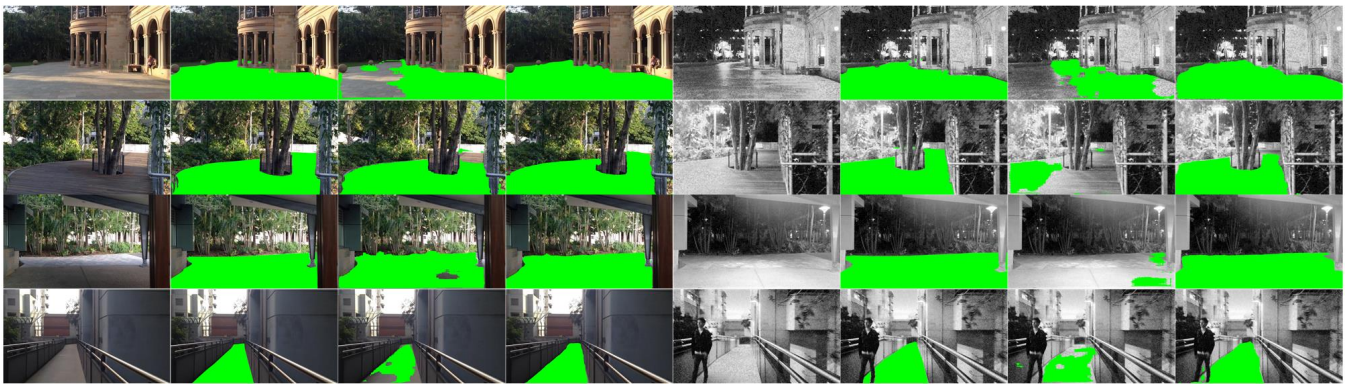
Fig. 3. Qualitative examples in the domain of panoramic annular semantic segmentation: (a) Raw panoramic images from the PASS dataset [2][14], (b) Unfolded panoramas, (c) Predictions of our ERF-PSPNet in the semantic spaces of Mapillary Vistas, (d) ADE20K and (e) IDD20K.

indicates that a ConvNet that needs to work in challenging unseen domains also needs to see challenging examples like cluttered scenes during training while data diversity should always be ensured. Note that all these results are obtained by viewing the panorama as a single input segment. In [2][14], it was shown that for panorama segmentation, partitioning the input into several segments and predicting with cross-segment padding are critical. In this work, our best results can also be significantly further improved by using the operations as highlighted in Table III. However, because there is not a teacher model that is significantly better than our jointly-trained efficient ConvNet on IDD20K dataset, we leave the omni-supervision for future work to particularly address omni-directional semantic segmentation.

Fig. 3 displays representative predictions of our approach in diverse semantic spaces. On the one hand, clear and robust segmentation in the unseen panoramic imagery can be observed. Besides, in this demonstration, it is shown that while only a single model is yielded, it delivers rich sets of visual classes and they are complementary to each other. For example, in the Vistas space, crosswalks and curbs can be predicted that are absent in IDD20K space (see the 3rd to 6th rows), but IDD20-space results can help to foresee safety-critical classes like auto-rickshaws whose behavior is highly unpredictable, as shown in the 4th/6th rows (denoted with yellow). Additionally, with ADE20K-space predictions, even indoor environments can be covered beyond IV's perception.

E. Multimodal Walking Domains

Detecting traversable area ahead of an intelligent robotic agent using an on-board camera is a key capability, which can benefit from pixel-wise semantic segmentation [1][7]. We use the Gardens Point dataset [15] to study the robustness in multimodal imagery. Gardens Point dataset has 600 images, of which 400 were captured at day and 200 were collected at night along nearly the same walking route across indoor and outdoor environments. The nighttime images were converted



(a) Day input (b) Annotation (c) ADE-trained (d) Omni-trained (e) Night input (f) Annotation (g) ADE-trained (h) Omni-trained

Fig. 4. Qualitative examples of traversable area segmentation across indoor and outdoor: (a) Day and (e) Night images from the Gardens Point dataset [15], (b)(f) Manually annotated ground truth, (c)(g) Predictions of ADE20K-trained and (d)(h) Omni-supervised ERF-PSPNet.

to grayscale with contrast enhanced as shown in Fig. 4. In this work, we manually create pixel-accurate annotations of traverasable areas for all the 600 images.

TABLE IV

ACCURACY ANALYSIS ON GARDENS POINT DATASET [15].

Network	Day	Night	All
ERF-PSPNet (Vistas)	75.2%	54.4%	68.4%
ERF-PSPNet (ADE20K)	78.8%	58.2%	71.7%
ERF-PSPNet (Jointly)	82.7%	66.7%	76.6%
Omni-Supervised	**83.7%**	**73.1%**	**80.0%**

Table IV shows the quantitative results of our omni-supervised solution contrasted with independently/jointly-trained methods after merging semantic traversable classes to facilitate comparison. It can be seen that ADE20K-based model is better than the Vistas-based model because it has seen both indoor and outdoor scenes in the training phase. The jointly-trained model significantly improves the scores as it learns more illumination-invariant features, becoming more accurate across RGB/grayscale modalities, indoor/outdoor and day/night scenarios. The omni-supervised proposal further improves the mIoU greatly, especially at night, reaching an overall accuracy of 80.0%. Fig. 4 also demonstrates that our omni-supervised approach consistently leads to more complete and robust segmentation across indoor and outdoor at both daytime and nighttime, beneficial for real-world navigation assistance systems.

F. Nighttime Driving Domains

IV applications can hardly escape from non-ideal or even adverse weather and illumination conditions. Since we have realized that our approach can lead to better performance at night, it is worthwhile to look into the segmentation for nighttime road-driving images. In this work, we use the Nighttime Driving Dataset [16] (50 testing images) and BDD Database [17] (32 nighttime images for validation). We compare qualitatively in each dataset both single-source Vistas-trained model and our omni-supervised model as it is not able to compare the numerical results due to the different classes hierarchies, but our approach already yields comprehensive semantics in multiple spaces. To analyze the robustness gains, as shown in Fig. 5, the background classes are better classified, the roadways are more completely/consistently segmented, while the dynamic participants are finely detected (see the trains). In all conditions, our proposal clearly exhibits considerable generalization benefits, enabling efficient ConvNets like ERF-PSPNet to work reliably even at night without any domain adaptation.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a multi-source omni-supervised learning framework to increase the number of recognizable visual classes and robustness of efficient architectures ready to be deployed in various domains. While previous researches that relied on regulating the class hierarchies also stated that efficient ConvNets cannot well handle the complexity, our approach enables them to deliver rich sets of detectable semantics, meanwhile improves their reliability in heterogeneous unseen domains. The high generalization capacity is empowered by multi-source data and data distillation by leveraging both labeled and unlabeled images, which expose the efficient ConvNet to diverse scenes.

The experimental results show that our solution with a single unified model ERF-PSPNet outperforms state-of-the-art efficient networks on IDD20K and PASS datasets, both of which reflect highly unconstrained road surroundings. We further investigate the performance in multimodal walking and nighttime driving domains, demonstrating consistent and significant robustness benefits across indoor and outdoor, even in adverse conditions such as the nighttime.

We have the intention to incorporate non-local operation and multi-domain adversarial learning [31] to further enhance the generalizability in the presence of unforeseen scenery. Particularly, we aim to optimize the omni-supervised solution for omni-directional semantic segmentation by importing unlabeled panoramas. In addition, while we consider the multi-space detection results are already very useful for upper-level applications, it remains future work to fuse the recognizable semantics in a single segmentation map.

REFERENCES

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

(a) Input (b) Vistas-trained (c) Omni-supervised (d) Input (e) Vistas-trained (f) Omni-supervised

Fig. 5. Qualitative examples of semantic segmentation in unseen nighttime domains: (a) Input images from nighttime driving dataset [16], (d) Input images from BDD dataset [17], (b)(e) Predictions of Vistas-trained and (c)(f) Omni-supervised ERF-PSPNet.

[2] 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," *arXiv preprint arXiv:1909.07721*, 2019.

[3] 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.

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

[5] 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.

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

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] M. Leonardi, D. Mazzini, and R. Schettini, "Training efficient semantic segmentation cnns on multiple datasets," in *International Conference on Image Analysis and Processing*. Springer, 2019, pp. 303–314.

[13] 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*. IEEE, 2018, pp. 4119–4128.

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

[15] N. Sünderhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, "On the performance of convnet features for place recognition," in *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2015, pp. 4297–4304.

[16] D. Dai and L. Van Gool, "Dark model adaptation: Semantic image segmentation from daytime to nighttime," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2018, pp. 3819–3824.

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

[18] 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.

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

[20] 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.

[21] 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*, 2018.

[22] M. Orsic, I. Kreso, P. Bevandic, and S. Segvic, "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*. IEEE, 2019, pp. 12 607–12 616.

[23] 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.

[24] 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.

[25] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille, "Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation," in *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2015, pp. 1742–1750.

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

[27] T. Kalluri, G. Varma, M. Chandraker, and C. Jawahar, "Universal semi-supervised semantic segmentation," in *2019 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2019, pp. 5259–5270.

[28] T.-H. Vu, H. Jain, M. Bucher, M. Cord, and P. Pérez, "Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2517–2526.

[29] 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," *arXiv preprint arXiv:1711.03213*, 2017.

[30] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in neural information processing systems*, 2017, pp. 1195–1204.

[31] A. Schoenauer-Sebag, L. Heinrich, M. Schoenauer, M. Sebag, L. F. Wu, and S. J. Altschuler, "Multi-domain adversarial learning," *arXiv preprint arXiv:1903.09239*, 2019.