International Journal of Computer Vision
https://doi.org/10.1007/s11263-024-02207-3

R^{2}S100K: Road-Region Segmentation Dataset for Semi-supervised Autonomous Driving in the Wild

Muhammad Atif Butt[1,2] · Hassan Ali[2,3] · Adnan Qayyum[2,4] · Waqas Sultani[2] · Ala Al-Fuqaha[5] · Junaid Qadir[6]

Received: 10 August 2023 / Accepted: 27 July 2024

Abstract

Semantic understanding of roadways is a key enabling factor for safe autonomous driving. However, existing autonomous driving datasets provide well-structured urban roads while ignoring unstructured roadways containing distress, potholes, water puddles, and various kinds of road patches i.e., earthen, gravel etc. To this end, we introduce Road Region Segmentation dataset (R^{2}S100K)—a large-scale dataset and benchmark for training and evaluation of road segmentation in aforementioned challenging unstructured roadways. R^{2}S100K comprises 100K images extracted from a large and diverse set of video sequences covering more than 1000 km of roadways. Out of these 100K privacy respecting images, 14,000 images have fine pixel-labeling of road regions, with 86,000 unlabeled images that can be leveraged through semi-supervised learning methods. Alongside, we present an Efficient Data Sampling based self-training framework to improve learning by leveraging unlabeled data. Our experimental results demonstrate that the proposed method significantly improves learning methods in generalizability and reduces the labeling cost for semantic segmentation tasks. Our benchmark will be publicly available to facilitate future research at https://r2s100k.github.io/.

Keywords Autonomous driving · Semantic segmentation · Semi-supervised learning

1 Introduction

Visual perception for recognizing objects, obstacles, and pedestrians is a core building block for efficient autonomous driving. Semantic segmentation has emerged as an efficient perception method that aims to determine the semantic labels for each pixel of an image (Siam et al., 2018). Thanks to the availability of rich scene segmentation datasets (discussed in Fig. 2), significant technical progress has been made in this direction. However, several formidable challenges still remain on the path to efficient autonomous driving in the wild.

Firstly, existing autonomous driving datasets (Brostow et al., 2009; Caesar et al., 2020; Cordts et al., 2016; Geiger et al., 2012; Sun et al., 2020; Yu et al., 2020) are not generalized; they cover well-paved urban roads of developed countries which represents 3.7% road infrastructure

Communicated by Hong Liu.

✉ Junaid Qadir
jqadir@qu.edu.qa

Muhammad Atif Butt
mabutt@cvc.uab.cat

Hassan Ali
hassan.ali@unsw.edu.au

Adnan Qayyum
adnan.qayyum@itu.edu.pk

Waqas Sultani
waqas.sultani@itu.edu.pk

Ala Al-Fuqaha
aalfuqaha@hbku.edu.qa

[1] Computer Vision Center (CVC), Universitat Autònoma de Barcelona Bellaterra, Cerdanyola del Vallès, Spain

[2] Information Technology University of the Punjab, Lahore, Pakistan

[3] University of New South Wales (UNSW), High St, Kensington NSW 2052, Sydney, Australia

[4] Qatar University, Doha, Qatar

[5] Information and Computing Technology (ICT) Division, College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar

[6] Department of Computer Science and Engineering, College of Engineering, Qatar University, Doha, Qatar

Published online: 23 August 2024

of the world (Schwab, 2019) and barely serve 17% of the total world's population (Gaigbe-Togbe et al., 2022). More recently, Segment Anything (Kirillov et al., 2023)—the largest segmentation dataset with more than one billion masks for 11 million images has been released to perform general purpose segmentation tasks. However, despite being the largest in size, it only covers 0.9% of data samples from low-income countries. Therefore, these datasets have scant coverage of unstructured roadways containing hazardous road patches (i.e., distress, earthen, gravel) that are common in the developing world, as shown in Fig. 1. Such ambiguous road regions pose an enormous hazard to human drivers and lead to severe road accidents and fatalities. According to World Health Organization (WHO), 1.3 million people die every year due to road accidents (WHO, 2020) with 93% of causalities occurring in low- and middle-income countries. The global road safety report points out that non-standard road infrastructure is a key reason for higher road accident rates in these countries (WHO, 2019). Therefore, the under-representation of such challenging data in existing datasets is a critical omission for research on autonomous driving and an indication of the need for a benchmark to improve autonomous driving in such challenging road scenarios.

Secondly, pixel-level annotation of images is excessively expensive—for cityscapes, labeling an image took an hour on average (Cordts et al., 2016)—leading to smaller segmentation datasets than in other domains (Deng et al., 2009; Lin et al., 2014), consequently limiting the generalizability of the trained models. Although semi-supervised learning methods (Abdalla et al., 2019; He et al., 2019; Huang et al., 2018; Yu et al., 2022) have been proposed that leverage unlabeled data to improve learning, these methods suffer limitations because (i) segmentation datasets are often highly imbalanced in terms of pixel counts corresponding to each class (Rezaei et al., 2020), and different physical scenarios in which the dataset is collected. Therefore, the resulting model performs significantly worse in physical scenarios that are not common (e.g. rare weather conditions and unstructured roads), which can be lethal in autonomous driving; (ii) Biased predictions caused by the data imbalance in early semi-supervised training phase (He et al., 2019) lead to a higher misclassification rate during inference; (iii) self-training segmentation models are computationally very expensive due to a large number of pseudo labels (Wei et al., 2018). In this regard, there is a need for an efficient method to improve performance while considering accuracy-energy trade-offs. To address these challenges, we have made the following contributions:

1. We introduce Road Region Segmentation (R^2S100K) dataset for autonomous driving comprising 100K diverse set of road images, covering 1000+ KMs of challenging roadways, as shown in Fig. 1. R^2S100K dataset covers more challenging road categories and scenarios than existing datasets. Moreover, R2S100K serves as an initial step in representing unstructured roads prevalent in low-income countries, allowing for a more comprehensive stress-testing of foundational segmentation models for autonomous driving.
2. We propose an unsupervised Efficient Data Sampling (EDS) method to sample a subset from the unlabelled training data, which offers three benefits: (i) EDS notably alleviates the data imbalance in the physical scenarios, (ii) improves the performance of supervised (0.71–6.72% MIoU) and semi-supervised (0.26–1.84% MIoU) models, and (iii) significantly reduces the annotation and training costs (75% fewer pseudo-labels and 79% decrease in the training time).
3. The EDS is compatible with multiple learning frameworks (supervised, semi-supervised) and model architectures. It can be integrated with datasets such as Cityscapes, CamVid, and BDD100K due to a similar labeling schema.

The rest of the paper is organized as follows. Section 2 presents the related work of autonomous driving benchmarks and datasets for 2D visual object detection and scene segmentation, and semi-supervised methods to perform the aforementioned tasks. Section 3 presents our proposed R^2S100K and the Efficient Data Sampling (EDS) enabled novel self-training method for drivable road region segmentation to distinguish safe and hazardous road regions. In Sect. 4, several state-of-the-art segmentation methods are evaluated to present a comprehensive benchmark study alongside the effectiveness of our EDS-based self-training settings. Lastly, concluding remarks are summarized in Sect. 6.

2 Background

2.1 Autonomous Driving Datasets

In the past couple of years, several datasets have been released to accelerate the development of visual perception algorithms. These datasets can be categorized into two major groups: (i) object detection—which focuses on 2D/3D objects (Caesar et al., 2020; Dollár et al., 2009; Geiger et al., 2012; Huang et al., 2019; Sun et al., 2020; Xiao et al., 2021; Zhang et al., 2017); and (ii) scene segmentation—which focuses on semantic segmentation for scene understanding. We present a detailed comparison of these state-of-the-art datasets in Fig. 2, highlighting their key attributes, such as image resolution, the number of images, and the diversity of regions and road types, while also emphasizing the unique

Fig. 1 Examples of our dataset images covering a wide array of roadways, varying across different lighting and weather conditions. Instead of considering the whole paved road region as one class, we distinguish safe asphalt road region and its associated atypical classes found on unstructured roads such as distress, wet surface, gravel, boggy, vegetation misc., crag-stone, road grime, drainage grate, earthen, water puddle, misc., speed breakers, and concrete road patches

differences between the state-of-the-art datasets alongside the comprehensive nature of our R2S100K dataset in terms of diversity and applicability to unstructured roadways. Here we discuss some important characteristics of these datasets. **Object Detection Datasets:** KITTI (Geiger et al., 2012) is one of the most widely used vision benchmark suites for object detection on urban roads and highways which contains 15k images along with 200k annotations. Later, Waymo open dataset (Sun et al., 2020) presented more than 23 million 2D and 3D bounding boxes annotations of 1150 inter-cities urban scene segments. The nuScenes (Caesar et al., 2020) dataset presented 1.4 million 3D bounding box annotations of 1000 urban and suburban road scenes for 23 classes. In 2019, ApolloScape dataset (Huang et al., 2019) has been released with comprises 70k 3D annotations along with 160k semantic mask annotations of urban roads and highways under varying weather conditions. Similarly, Pandaset (Xiao et al., 2021) presented 1 million 3D bounding box annotations for object detection in urban traffic scenarios. Other than these, various other datasets (Caesar et al., 2020; Dollár et al., 2009; Zhang et al., 2017) have been proposed, which played an important

Dataset	Images	Resolution	No. of Cities
KITTI	400	1242 x 375	1
CamVid	700	672 x 453	1
CARL-D	7,500	1920 x 1080	50
IDD	10,003	1920 x 1080	N/A
Cityscapes	25,000	2048 x 1024	27
A2D2	48,000	1920 x 1280	1
BDD100K	100,000	1280 x 720	4
nuScenes	1.4 M	1600 x 900	2
Waymo Open	1 M	1920 x 1280	3
MVD	25,000	1920 x 1280	N/A
WD2	4,256	1920 x 1280	N/A
R²S100K	**100,000**	**1920 x 1080**	**12**

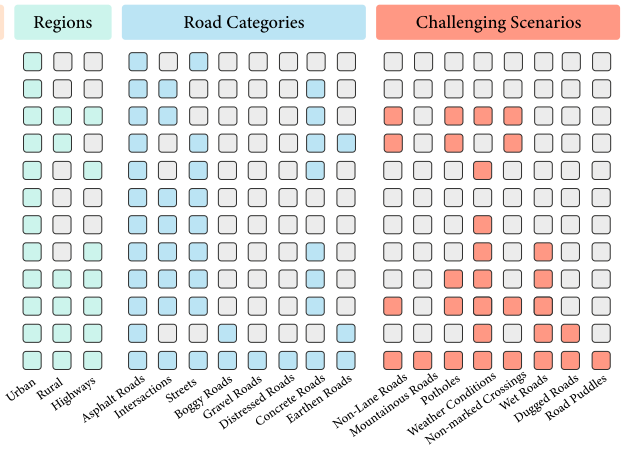

Fig. 2 Comparison of dataset statistics with existing driving datasets i.e., KITTI (Geiger et al., 2012), CamVid (Brostow et al., 2009), CARL-D (Butt & Riaz, 2022), IDD (Varma et al., 2019), Cityscapes (Cordts et al., 2016), A2D2 (Geyer et al., 2020), BDD100K (Yu et al., 2020), nuScenes (Caesar et al., 2020), Waymo (Sun et al., 2020), MVD (Neuhold et al., 2017), and Wilddash (Zendel et al., 2018). Our R^2S100K covers more diverse road infrastructure and challenging scenarios than the existing benchmarks. Therefore, our dataset can be used to develop more robust and generalized road segmentation methods for autonomous driving

role in developing efficient object detection and recognition algorithms.

Semantic Segmentation Datasets: CamVid (Brostow et al., 2009) is considered among the pioneer scene segmentation datasets—comprising 700 fine annotations for 32 classes. In 2016, Cityscapes (Cordts et al., 2016) was released, which contains 5000 fine and 20,000 coarse annotations for urban roads. In 2017, Mapillary Dataset (Neuhold et al., 2017) comprising 25K fine annotations of inter-continental urban scenes was presented. Later on, BDD100K (Yu et al., 2020) is released in 2020, which provides 10K fine annotations of urban roadways. MVD (Neuhold et al., 2017) contains 25K images covering diverse yet urban roadways.

Though these datasets provide enriched information on urban scenes for scene segmentation tasks, they do not cover unstructured road conditions and hazardous road patches, commonly encountered in developing countries. Therefore, models trained on these datasets cannot be generalized to the challenging roadways. Besides urban driving, a few datasets have been released for visual perception in off-road driving scenarios. OFFSEG (Viswanath et al., 2021) framework covers RELLIS-3D (Jiang et al., 2021) containing 6235 images, and RUGD (Wigness et al., 2019) comprising of 7546 images of outdoor off-road driving scenes. Wilddash-v2 contains 4256 images (Zendel et al., 2018) and covers unstructured road classes like distress and gravel patches. However, they label these classes under single *Road* class rather than distinguishing them as safe and hazardous regions. Recently, CARL-D (Butt & Riaz, 2022; Rasib et al., 2021), and IDD (Varma et al., 2019) datasets have also been released which provide annotations of urban and rural roads, however, they still lack aforementioned hazardous road patches that can highly influence the performance of autonomous driving models.

2.2 Scene Segmentation Methods

Fully Supervised Learning: Since the pioneering work of FCN (Long et al., 2015), significant progress has been made in developing deeper neural networks for semantic segmentation tasks. The semantic segmentation model aims to predict the semantic category of each pixel from a given label set and segment the input image according to semantic information—suggested by Long et al. (2015). The FCN outperforms conventional approaches by 20% on the Pascal VOC dataset. The U-net is an idea by Ronneberger et al. (2015) for segmenting biological images. U-net has a spatial path to maintain spatial information and a context path to learn context knowledge.

Later, various supervised methods (Badrinarayanan et al., 2017; Chen et al., 2014, 2017a, b, 2018; Noh et al., 2015; Romera et al., 2017; Yu et al., 2018; Zhao et al., 2017, 2018; Zhang et al., 2022) have been proposed to perform segmentation tasks efficiently. However, these methods employ deep CNNs as backbone networks, which require an immense amount of time to annotate large-scale data, limiting the model's capacity to adapt and further improve segmentation performance.

Semi supervised Learning: Recently, semi-supervised learning methods have demonstrated better applicability in several segmentation domains. These methods have achieved state-of-the-art performance on several segmentation tasks by leveraging a huge amount of unlabeled data. In literature, several techniques such as video label propagation (Budvytis

et al., 2017; Luc et al., 2017; Mustikovela et al., 2016), knowledge distillation (Xie et al., 2018; Liu et al., 2020), adversarial learning (Huang et al., 2018; Souly et al., 2017), and consistency regularization (Mittal et al., 2019) are employed to perform semi-supervised segmentation.

Recently, (Chen et al., 2021a) proposed a consistency regularization method named Cross Pseudo Supervision, which enforces consistency between two perturbed networks with different initialization, effectively expanding training data using unlabeled data with pseudo labels. In another research work, Ouali et al. (2020) proposed a cross-consistency-based semi-supervised training method that enforces consistency between the main decoder predictions and auxiliary decoders' outputs that ultimately enhances the encoder's representations and thus leads to the improved results on SOTA datasets.

3 Methodology

In this section, we describe the R^2S100K dataset and our proposed efficient self-training method for semantic segmentation tasks. Figure 2 compares our dataset with existing datasets. This section introduces a benchmark suite for our proposed Road Region Segmentation Dataset (R^2S100K). Firstly, we describe R^2S100K in terms of the methodology adopted for data collection, frame selection, labeling, and distribution. Secondly, we discuss the categorization of supervised/ semi-supervised learning methods to develop a benchmark suite for our proposed dataset. In the later section, we discuss our proposed EDS-enabled teacher-student-based efficient self-training approach to solving the data imbalance problem for semantic segmentation tasks.

3.1 R^2S100K

We present a large-scale R^2S100K dataset to train and evaluate supervised/semi-supervised methods in challenging road scenarios. Our dataset can be distinguished from existing datasets in the following three major aspects:

Distribution Shift: R^2S100K dataset covers unique and undesiring urban and rural road conditions—described in Table 1 which are commonly encountered while driving, especially in developing countries. Whereas, existing datasets such as KITTI (Geiger et al., 2012), CamVid (Brostow et al., 2009), Cityscapes (Cordts et al., 2016), A2D2 (Geyer et al., 2020), MVD (Neuhold et al., 2017), BDD100K (Yu et al., 2020), nuScenes (Caesar et al., 2020), Waymo (Sun et al., 2020) represent well developed urban roadways, as depicted in Fig. 3. IDD though covers distressed and muddy road regions, however, it only distinguishes the mud class from the road and covers damaged road patches under one road class. Moreover, OFFSEG (Viswanath et al., 2021) The framework primarily covers off-road driving scenes, which significantly differ from unstructured roadways regarding representation. Similarly, Wildash (Zendel et al., 2018) covers distress and gravel patches under a single *Road* class rather than distinguishing them as safe and hazardous regions.

Diversity: R^2S100K is constructed over road sequences—captured from 1000+ KMs roadways of Pakistan considering diverse terrain, infrastructural features, and environmental attributes as shown in Fig. 4. To ensure diversity in data, we primarily focus on including motorways, highways, and urban traffic roads from Punjab, the largest province of Pakistan in terms of population (approximately 127.474 million). Additionally, we extend our coverage to encompass the rural and hilly areas of Khyber-Pakhtunkhwa, the second-largest province by population (approximately 35.53 million), operating under diverse illumination and weather conditions.

Generalizability: R^2S100K covers a diverse range of road infrastructure, including well-paved asphalt roads along with associated unique hazardous road regions which are categorized as atypical classes, enlisted in Table 1. However, we assigned distinct labels for our anomalous road classes and used similar labeling schema for asphalt class as cityscapes and BDD100K to ensure the integration of datasets for domain adaptation and semi-supervised learning.

3.1.1 Data Acquisition

Driving Platform Setup: A camera is mounted over the dashboard of a standard van with a height of 1.4 m from the ground and configured to an aspect ratio of 16:9 to capture the ultimate width of the road. A camera stabilizer is also installed to reduce the vibration effects of the vehicle.

Road Video Collection: We carefully followed the travel advisory issued by the government to identify diverse roadways. Based on the analysis, we defined a route plan to cover diverse infrastructure for data collection (as shown in Fig. 4) to ensure the inclusion of highways, expressways, and general roads of urban cities, rural and hilly areas.

Data Quality Control: We performed pre- and post-collection quality control (QC) to ensure high-quality data collection. In pre-collection QC, the data engineer must set up and monitor the camera's data stream while recording. Post-collection QC required data engineers to manually identify and remove the distorted/over-exposed/unclear video sequences. In our post-collection QC process, our data engineers meticulously apply multi-step checks to identify and exclude distorted, blurry, and unclear sequences from our dataset. Our quality check criteria encompass various factors, including but not limited to:

Ensuring Clarity: Firstly, we assess the sharpness of images using Structural Similarity Index and Gradient Magnitude to measure the sharpness of individual frames within each sequence quantitatively. Frames with low sharpness scores,

Fig. 3 Examples of road types covered in existing autonomous driving datasets for visual scene segmentation. R2S100K covers more challenging/hazardous roads in both—the urban and rural areas. While most of the existing datasets focus on the well-paved road infrastructure of urban areas and do not distinguish between safe and hazardous road region

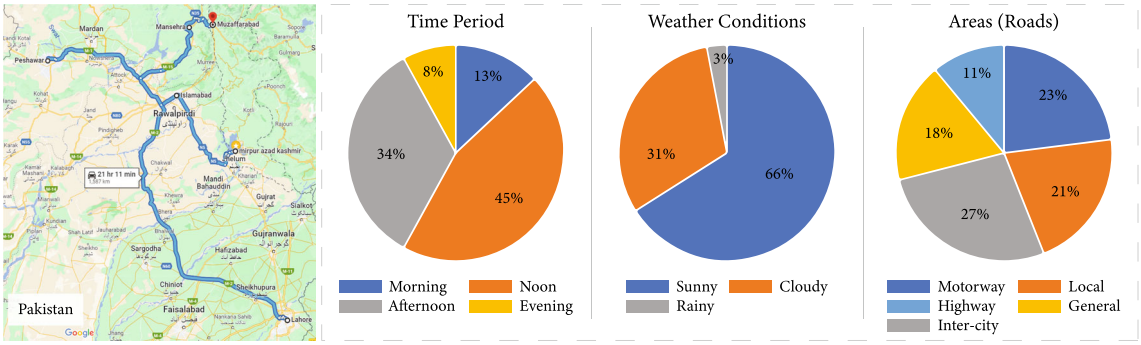

Fig. 4 Statistical analysis demonstrating the diversity of R^2S100K Dataset. (Left) Google Maps of routes covered for data collection. (Right) Different environmental and infrastructural characteristics: (1) timestamp, (2) weather conditions, and (3) road hierarchy. We cover over 1000 KMs of roadways of Pakistan—carefully considering the inclusion of motorways, highways, general inter-city and intra-city roads, as well as the rural and hilly areas, under different illumination and weather conditions

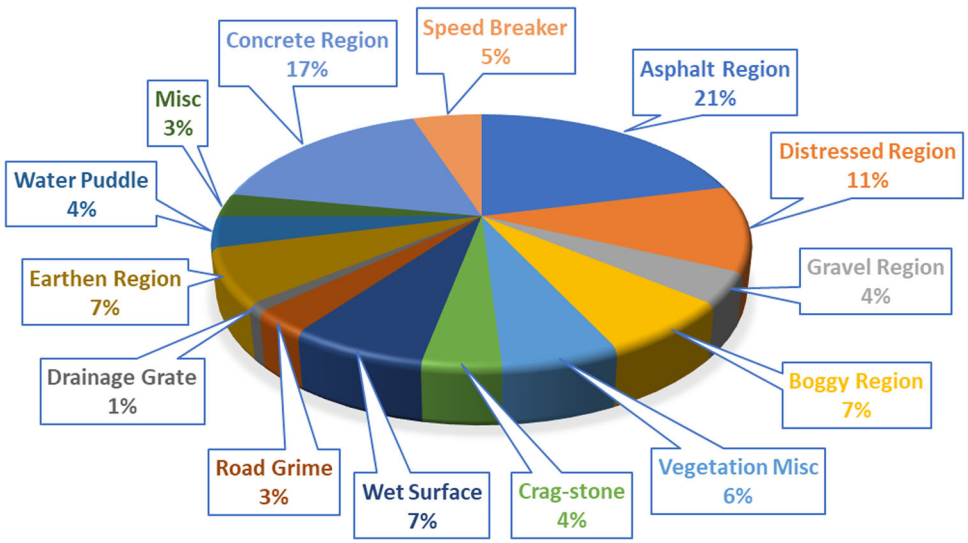

Fig. 5 Distribution of road classes in R^2S100K. Asphalt and concrete regions represent the safe drivable road regions with the higher representation among the other hazardous road patches

Table 1 List of classes along with their definitions

Class	Definition
Asphalt	Road pavement constructed using aggregates (crushed rocks, sand, and coal tar)
Distress	Longitudinal and transverse cracks occurred due to lack of maintenance
Gravel	Unpaved surface with loose aggregation of variable-sized fragments of rocks
Boggy	Unpaved road surface filled with mud
Vegetation Misc	Naturally occurring vegetation (other than trees) adjacent to the road
Crag-stone	Hilltop stones—dropped over road surface in mountainous areas
Wet Surface	Slightly watered road surface; can be damp due to snow or cold weather
Road Grime	Dirt ingrained on the road
Drainage Grate	An elongated cover with holes in it or a grating used to cover a water drain
Earthen	Unpaved roads with compacted layers of stabilized soil
Water Puddle	Small pool of water over the road
Misc	An unclear object dropped over the road
Concrete	Binders such as rough and fine aggregates
Speed Breaker	Concrete speed bumps, speed humps, and speed cushions over the road surface

indicative of blurriness, are flagged for visual/manual evaluation.

Detecting Distortion and Blur: Secondly, we analyze the contrast and exposure levels of frames to identify instances of distortion or motion blur. Histogram analysis is utilized to evaluate the distribution of pixel intensities and detect anomalies related to over-exposure or under-exposure.

Assessing Relevance and Consistency: We prioritize frames that best represent diverse road conditions and scenarios to ensure the relevance and representativeness of our dataset while striving for uniformity among the images to maintain consistency across the dataset. Our team conducted rigorous visual inspections of each video sequence. Trained evaluators assessed the overall clarity, distortion, and visual fidelity of the frames, considering factors such as motion blur, lens aberrations, and compression artifacts.

Data Distribution: After data collection under different illumination and weather conditions from 1000+ KM of roadways, distorted/blurry/unclear sequences are excluded, and frames are selected from the remaining video with a 10 s difference to avoid redundancy. The vehicle is moving at varying speeds [120 km/h (motorway), 60–100 km/h (highway), 20–60 km/h (within city)]. Therefore, speed variation, blurry sequence exclusion, and 10 s difference are key to avoiding data redundancy. Lastly, EDS further minimizes the chances of sequential frames in the data. We aligned video sequences to extract the frames to distribute the diverse road scenarios equally. To achieve better diversity, 10 frames are selected after every 10 s per frame. Therefore, 100K images of R^2S100K dataset are sampled out of 10 million images.

3.1.2 Data Statistics

Labeled Data: The labeled set consists of 14,700 images with fine-layered polygonal annotations which are realized in-house to ensure the highest level of quality. Firstly, annotators were provided with extensive training sessions to familiarize them with the data categorization, classes, and annotation tool to ensure consistency and accuracy. During training, similar data samples were distributed to the data annotators to allow for cross-verification, and the labeling strategy has been refined through iterative Inter-Annotator Agreement considering the definitions of the road classes. Secondly, to avoid void spacing and erroneous class overlapping, images are labeled back to front so that no class boundary is dual-labeled. Due to the diversity in data, we categorized road regions into 14 distinct classes as described in Table 1. Additionally, to further facilitate the annotators, we use (SuperAnnotate AI Inc., 2024) for labeling which is a user-friendly tool especially for autonomous driving tasks. In the post annotation phase, a random sampling and expert validation has been performed by the experts to cross-evaluate the quality of annotations, and to identify and address the errors, ensuring the correctness and reliability of R2S100K dataset.

Unlabeled Data: The unlabeled set of our dataset contains 86,000 images, covering diverse road infrastructure. As shown in Fig. 4, our unlabeled set is collected under varying weather conditions and time periods to ensure diversity regarding downstream autonomous driving tasks.

3.2 Training Fully Supervised Baseline Models

To analyze the effectiveness of R^2S100K, we fine-tuned SoTA segmentation networks to leverage the representations from pre-trained weights learned from large-scale datasets, enhancing the generalizability of models to our road region segmentation tasks. These models include FCN (Long et al., 2015), PSPNet (Zhao et al., 2017), FPN (Lin et al., 2017), LinkNet (Chaurasia & Culurciello, 2017), Deeplabv3+ (Chen et al., 2018), and LRASPP (Howard et al., 2019), MaskFormer (Cheng et al., 2021b), and SegFromer (Xie et al., 2021) along with various backbone networks to perform road region segmentation tasks. These methods are trained using a set of human-labeled images (x, y) where $x \in R^{H\times W\times 3}$ is a 3-channel RGB image, and $y \in R^{H\times W\times C}$ is a respective segmentation mask where H and W refers to height and width of the mask, and C indicate classes present in that mask. Following common practices

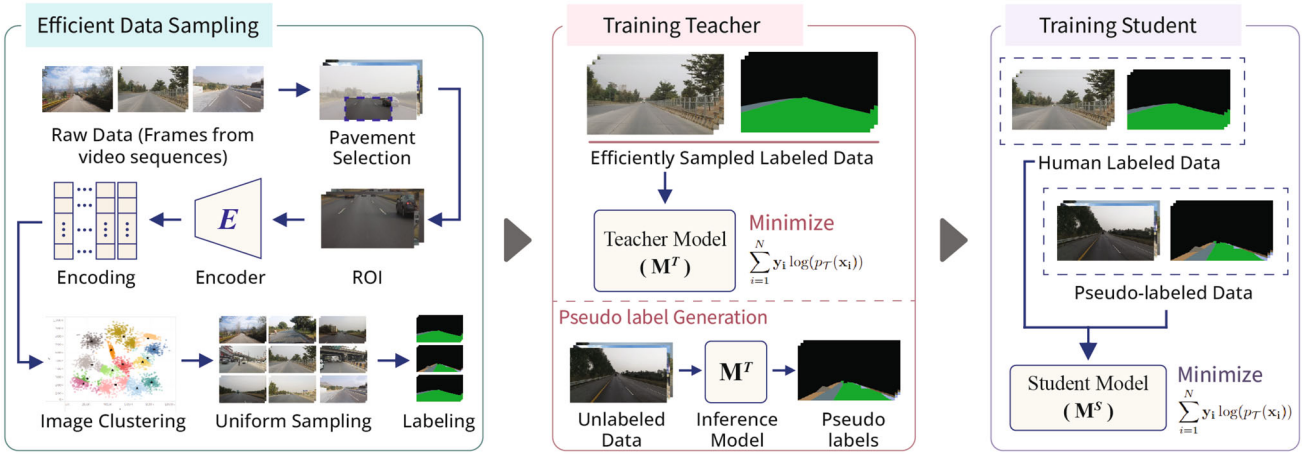

Fig. 6 Our Efficient Data Sampling (EDS) based self-training framework. Firstly, raw data samples are clustered based on similarity in road classes among image encodings (shown in Fig. 7)—generated by an encoder. Then, a small subset is uniformly formed from all clusters for annotation to train the teacher model. After training, pseudo-labels of the unlabeled set are generated using the teacher model, and the student model is trained on real and pseudo-labeled sets to achieve better generalization

Fig. 7 Visualizing examples of clusters (twelve clusters representing three images in each) using our EDS. Our EDS efficiently clusters images concerning the similarities in road texture, luminous conditions, and road scenarios

(Zhu et al., 2019), model M is trained using cross-entropy loss, and IoU is used as a performance metric.

3.3 Improving Self-Training Using Unlabeled Data

Recently, a surge of interest has been observed in utilizing unlabeled data to scale up the adaptation of deep models in various segmentation tasks. Leveraging many unlabeled sets from our R^2S100K, we carefully employ semi-supervised

training methods to study the generalizability of these models. We take inspiration from Zhu et al. (2019) and employ a teacher-student-based self-training framework for road segmentation. The student-teacher framework offers a structured approach to transfer rich representations and intricate spatial relationships from the teacher to the student. This guidance is particularly beneficial in tasks like road region segmentation, where precise delineation of spatial boundaries is crucial. Unlike directly using a pre-trained CNN/transformer model, which may overlook the nuanced insights captured by the teacher, the teacher-student framework facilitates focused knowledge transfer, leading to improved performance and more accurate segmentation results in complex real-world road scenarios.

Teacher-student-based self-training refers to an approach in which a large DL model (called the teacher) is trained using real labeled data. Then, a set of unlabeled images is given as input to the trained teacher model for inference, and the teacher model's output is considered a pseudo-label for the corresponding input image. Finally, data with both—the real and pseudo-labels are combined to train a small/different DL model (called student model) to learn representations from whole data. The purpose of training the teacher model on real data is to guarantee its performance in generating pseudo labels. Therefore, we utilize a small labeled set along with a large unlabeled set to increase the accuracy of the trained model while mitigating the human effort in producing labels at scale. Similar to the practices in supervised learning, we fine-tuned these models to leverage the already learned representations from large-scale datasets for faster convergence.

3.3.1 Efficient Data Sampling (EDS)

In semi-supervised segmentation, dealing with the data imbalance problem is highly challenging. In street scene segmentation problem, two key factors cause data imbalance; (i) *class imbalance*, which includes class-wise pixel imbalance—a typical image is largely occupied by sky and road, while other classes like humans and bicycles represent far fewer pixels—and class object confusion—some classes, e.g., bicycles, are more challenging to segment due to their complex shapes, occlusions, and faded representations (Brostow et al., 2009; Cordts et al., 2016); and (ii) *imbalance in physical scenarios*, as highlighted in Fig. 4. Although both imbalances are equally important to address, class imbalance is a post-annotation issue that mainly depends on the underlying task, and is generally easily detected, e.g., by computing the confusion matrix of each class. On the contrary, an imbalance in physical scenarios is a pre-annotation issue inherent to the (unlabeled) images. Further, physical scenarios under-represented in the training set are also usually equally under-represented in the test set. Thus, detecting imbalances in physical scenarios is significantly challenging, let alone alleviating them. We identify a dire need for an efficient method to detect/alleviate data imbalances in physical scenarios at the pre-annotation stage to produce more balanced models on semantic segmentation tasks. The Fig. 8 shows the KL-divergence from the uniform distribution of physical scenarios in two subsets from the original dataset - (i) the randomly sampled subset; and (ii) the EDS sampled subset. Ideally, the sampled subset should represent different physical scenarios equally, resulting in a uniform distribution. For example, in the sampled datasets, all the times (Morning, Noon, Afternoon, and Evening) should be uniformly represented. Therefore, a lower KL divergence of the sampled subset from the uniform distribution indicates a better sampling strategy. EDS notably improves the imbalance in different physical scenarios, as illustrated by the reduced KL-divergence in Fig. 8, and consistently better performance of the models in Fig. 11, respectively.

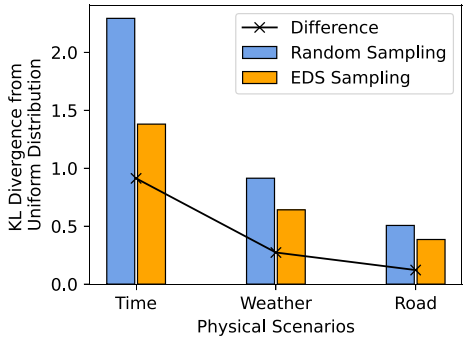

Fig. 8 KL divergence between both—the EDS and Random sampling-based data distributions

To address these issues, we propose EDS, as depicted in Fig. 6. We aim to equally represent different physical scenarios in the training data. In this regard, our EDS approach has two main stages: (i) data categorization, and (ii) data selection.

Data Categorization: Firstly, given an unlabeled dataset, $\mathcal{D}_x$, for each $x \in \mathcal{D}_x$, we extract region-of-interest (ROI) mainly comprising salient road features, sidewalks, and pedestrians, while ignoring background, e.g. sky. The extracted image ROI(x) is then processed through an off-the-shelf encoder network $e(\cdot)$ to get encodings $e(\text{ROI}(x))$. We use a U-Net model, built upon VGG-16 Imagenet encoder, $e : \mathcal{R}^{512\times512\times3} \rightarrow \mathcal{R}^{32\times32\times512}$, as backbone. Due to the prevalent data imbalance problem in segmentation datasets, inherent biases in datasets are also reflected in trained models. Whereas, models trained on Imagenet learn more generic features spanning over 1000 classes, and can be used for multiple downstream tasks. We feel using a biased encoder (trained on street scene dataset) in EDS to mitigate biases in R^2S100K is counter-intuitive.

Data Selection: Secondly, encodings $e(\text{ROI}(x))$ of unlabeled train set are fed to k-means to get k data clusters $\{C_i\}_{i=1}^{k}$

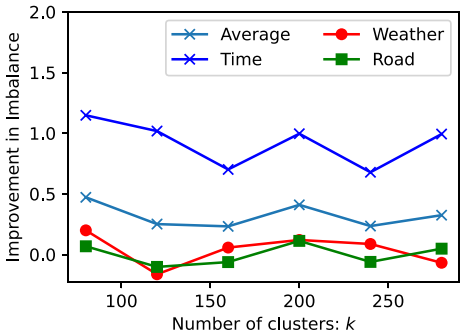

Fig. 9 Demonstrating the effect of change in k on data imbalance regarding physical scenarios

based on similarities in road surface. Finally, to maintain equal distribution along all types of road representations, we uniformly sample n data instances from each cluster, C_i, so that our final dataset, $\mathcal{D}_x^*$ has $n \times k$ data samples. In typical settings, we choose $n \times k = 3000$ to have a comparable dataset size as the Cityscapes dataset. Formally,

$$\mathcal{D}_x^* = \cup_{i=1}^{k} \{x_j \sim C_i\}_{j=1}^{n} \tag{1}$$

We choose k = 15 × 20 = 300. Although k is a hyperparameter, our goal for choosing k = 300 is to allow each of the 15 classes to be captured in 2(sun/no sun) × 2(rain/no rain) × 5(road areas) = 20 clusters representing different scenarios. In Fig. 9, we demonstrate the analysis regarding the influence of change in K on the average sampling. It can be seen that the change in K does not influence the sampling pattern. Therefore, setting k = 300 ensures the balanced sampling of physical road scenarios while keeping weather conditions and road classes in view. To compare EDS with random sampling, we sample 500 images from the original dataset using each method and compute the probability density of each physical scenario in Fig. 8 based on two sampled subsets. Ideally, all labels should have a uniform density, signifying equal representation in the dataset. Therefore, we compute KL-divergence between probability density and uniform distribution in Fig. 8. Results show that EDS significantly improves data imbalance as compared to random sampling.

3.3.2 Student–Teacher Method for Segmentation Task

Our self training framework is illustrated in Fig. 6. Based on better performance in supervised learning, bestperforming model is selected as teacher model T which is used to generate pseudo labels of our unlabeled set of images. The teacher model is used to generate pseudo labels y of our unlabeled set of images x. Similar to supervised learning, one-hot encoding of the class labels is sampled from the $p_{\mathcal{T}}(x)$ as given in equation.

$$L_{\mathcal{T}} = -\sum_{i=1}^{N} \mathbf{y}_i \log(p_{\mathcal{T}}(\mathbf{x}_i)), \tag{2}$$

where N denotes the number of labeled samples. y_i is the one-hot encoding of class labels, while $p_{\mathcal{T}}$ represents softmax predictions from the teacher model containing class probabilities.

We demonstrate various examples of our teacher-generated pseudo labels in Fig. 10. Thanks to our well-performing teacher model, the quality of our teacher-generated pseudo labels x over the unlabeled set is closer to human-annotated labels despite a large domain gap. Therefore, we combine pseudo and real labeled sets to train the student model S. Therefore, we combined pseudo and real labeled sets to train the student model S. Thanks to the generalizability of our proposed self-training pipeline, any DL-based segmentation model can be used as a student model irrespective of their network architectures (briefly explained In Sect. 4.6). Following the practice—adopted in supervised learning, the focus is set to minimize the cross-entropy, given in Eq. 3.

$$L_S = -\sum_{i=1}^{N} \mathbf{y_i} \log(p_{\mathcal{T}}(\mathbf{x_i})) - \sum_{j=1}^{M} \mathbf{y_j'} \log(p_{\mathcal{S}}(\mathbf{x_j'})) \tag{3}$$

M denotes the number of unlabeled samples. p_S represents softmax predictions from the student model containing the class probabilities. The predicted class probabilities of the student model will be near one-hot by training on hard pseudo-labels generated by the teacher model. Therefore, the entropy of unlabeled data is minimized with cross-entropy loss.

It is worth noting that the teacher model may generate noisy or incorrect pseudo-labels against rare/challenging scenes, which can significantly impact the training process of student models and ultimately hinder the overall performance. Therefore, we adopted a feedback-based training and evaluation approach to achieve maximum accuracy. The student model is first trained with real and pseudo-labeled data and then evaluated on the real validation set, which is common for both the teacher and student models. In the second step, the data engineers perform error analysis based on the IoU and confidence thresholding on the validation set to identify the source of misclassification. After completing the error analysis, the training set is regularized using EDS, where data samples of the identified class are augmented to improve convergence.

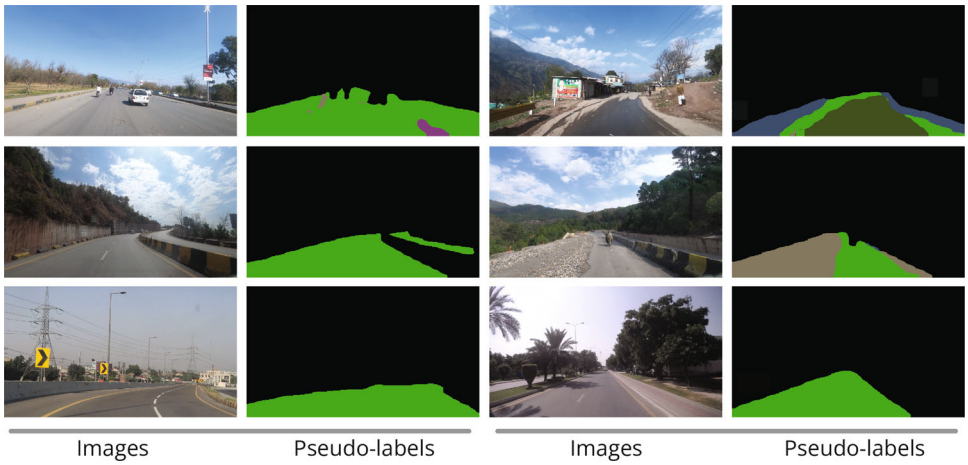

Fig. 10 Demonstration of our teacher-generation pseudo-labels over diverse roads. Our teacher model can provide reasonable segmentation predictions

Table 2 Evaluation of baseline segmentation methods by training using different numbers of randomly sampled sets from the actual train set of R^2S100K dataset for supervised learning

Model	Backbone	MIoU				
		1K	3K	5K	7K	9K
FCN	ResNet-101	41.07	54.48	54.21	54.02	53.62
PSPNet	ResNet-101	39.83	53.03	52.96	52.43	52.14
LRASPP	MobileNet-v3	36.31	56.54	56.19	56.10	55.93
FPN	ResNet-101	44.26	55.65	54.27	54.23	54.18
LinkNet	ResNet-101	43.50	56.14	55.71	55.06	54.84
SegFormer	–	49.77	57.86	57.60	57.48	57.21
MaskFormer	–	51.35	57.98	56.37	57.72	57.05
Deeplab-v3+	ResNet-101	45.97	**58.02**	57.36	55.89	55.37

4 Experiments and Results

Firstly, we briefly describe the implementation details regarding hyper-parameter selection for training and evaluating supervised and semi-supervised learning methods. We categorize our experiments into five sections. In Sect. 4.1, we analyze the performance of supervised learning methods and compare the results between random data sampling and our proposed EDS method. Section 4.2 evaluates the performance of semi-supervised learning-based standard self-training methods leveraging our unlabeled data. In Sect. 4.3, we select the best-performing semi-supervised model as the teacher method and evaluate its efficacy of the student model with different ratios of unlabeled data samples. In Sect. 4.4, we analyze the generalization of other student models irrespective of different network architectures. Lastly, we evaluate the cross-domain generalization with the same categories on state-of-the-art autonomous driving datasets, including Cityscapes, CamVid, IDD, and CARL-D.

4.1 Basic Settings

Following the training practices from the Cityscapes and BDD100K, the learning rate is set to 0.0001 for fine-tuning with SGD as an optimizer. As per conventional practice (Liu et al., 2015), a polynomial learning rate is used to smooth learning, and batch size, momentum, and weight decay are set to 8, 0.9, and 0.0001, respectively. Nvidia RTX 3060 is used to perform experiments. The number of training epochs is set to 200 with validation patience of 10 epochs. In R2S100K, each class is divided into three portions: 60% for training, 30% for validation, and 10% for testing to ensure a balanced representation of classes across the training, validation, and test sets, facilitating fair model evaluation and performance comparison. Evaluation is done using pixel accuracy, precision, recall, F1-score, and standard Jaccard index (shown in Eq. 4), where FP, TP, and FN refer to the number of false positive, true positive, and false negative pixels, determined over the test set.

$$\text{IoU} = \frac{TP}{(TP + FP + TN)} \tag{4}$$

4.2 Performance of Supervised Learning with EDS

We employed FCN (Long et al., 2015), PSPNet (Zhao et al., 2017), FPN (Lin et al., 2017), LinkNet (Chaurasia & Culurciello, 2017), Deeplabv3+ (Chen et al., 2018), LRASPP (Howard et al., 2019) MaskFormer (Cheng et al., 2021b),

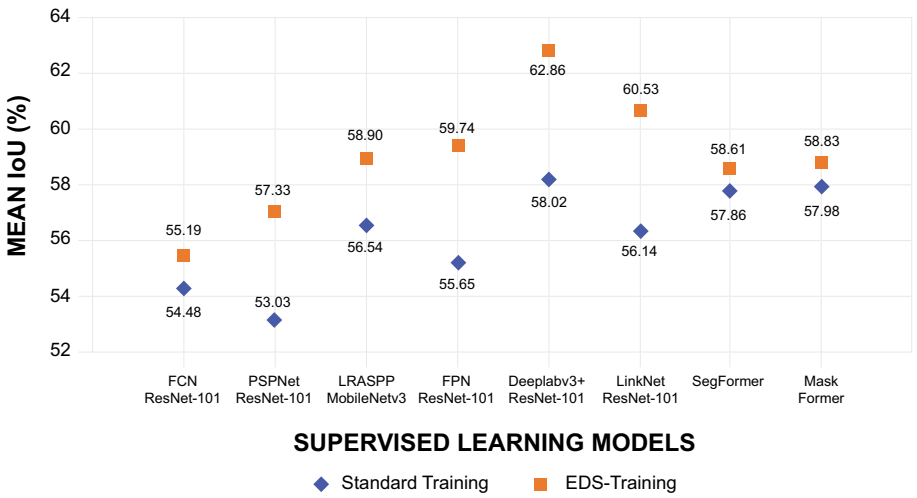

Fig. 11 Comparative analysis of baseline segmentation methods using standard data sampling (STDS) and our EDS. Our efficient data sampling method significantly improves supervised learning for semantic segmentation tasks

and SegFormer (Xie et al., 2021) with various backbones on R^2S100K. We analyze baseline segmentation methods over many labeled subsets (1k, 3k, 5k, 7k, and 9k), randomly sampled from actual 9000 train images. The primary motivation for adopting this training scheme is that in autonomous driving, the cameras are mounted in the center-straight orientation on the front (similar in Cityscapes, KITTI, BDD100K, IDD, and CARL-D) to capture the frontal view. Resultantly, the data, either in the form of video or frames, is sequential in nature with the static frame structure, i.e., the road in the lower-center, buildings, or trees on the left and right, while the sky covers the top-center part of the image. Secondly, the major part of the road is the asphalt and concrete region with the chunks of the other hazardous classes, defined in Table 1, and demonstrated in Fig. 1. In Fig. 5, it can be observed that asphalt and concrete road regions cover 39% of the labeled pixels in the dataset, which leads to highly unbalanced data. Considering these important factors, we hypothesize that using such unbalanced data may cause overfitting, ultimately hindering the model's generalization performance. From Table 2, it can be seen that employed models—trained over 1k images experience the worst performance due to under-fitting. However, their performance significantly improves with the 3K train set. Interestingly, the employed models start saturating while training on large train sets, i.e., 5k, 7k, and 9k samples, and do not further improve learning because of the similarity in road pavement across training samples.

We further analyze the performance of employed models using two data sampling methods, i.e., the standard training data selection (STDS) method—in which the data samples are randomly selected based on their occurrence, and our proposed EDS method. It is clear from Table 2 that segmentation methods perform well with a 3k training set. Therefore, we randomly select 3000 labeled images from the train set in STDS based on the frequent occurrence. On the other hand, using our EDS, we first clustered all images based on their representation similarities, shown in Fig. 6. Then, we uniformly sampled 3000 labeled images from all clusters to form a representative sub-training set.

The results illustrated in Fig. 11 show that our EDS method significantly improves learning in segmentation tasks. For instance, Deeplab-v3 with ResNet-101 achieved a comparatively highest mIoU, i.e., 62.86% using our EDS method which is 6.72% higher than its baseline trained using the STDS method. A major reason for this performance increase is that most of the informative data samples are ignored during random selection, due to which, training data becomes highly unbalanced which ultimately leads to inefficient training and poor generalization. Consequently, the resultant model does not achieve better performance on test data. In our EDS method, training samples are uniformly selected based on their class representations. Therefore, the network efficiently learns an equal distribution of features from each class, boosting the performance of trained models over test data. A class-wise comparison of state-of-the-art segmentation models is shown in Table 3.

4.3 Effectiveness of Student–Teacher Self-Training

Based on higher performance in supervised learning, we select DeepLabV3+ with ResNet101 as a teacher model to initiate self-training. Firstly, we generate pseudo labels of an unlabeled set with several subsets, as shown in Table 4. Then, a student model, i.e., PSPNet, is trained on real and pseudo-labeled sets. From Table 4, it can be observed that utilizing pseudo labels significantly improves segmentation models, which indicates that segmentation models can be improved using pseudo labels without large-scale labeled data.

Table 3 Segmentation results (in percentage) of baseline fully-supervised models using EDS on our R^2S100K dataset

Classes	Methods							
	FCN w/ ResNet-101	DeeplabV2 w/ ResNet-101	FPN w/ ResNet-101	FarSeg w/ ResNeXt-50	ICNet w/ ResNeXt-50	FastSCNN w/ ResNet-50	HR-Net w/ ResNeXt-101	PAN w/ ResNeXt-101
Asphalt Region	74.20	74.31	78.08	81.72	84.45	87.97	87.86	72.02
Wet Surface	66.76	69.59	67.58	69.13	70.10	71.47	70.36	68.87
Distress Region	58.02	61.48	59.74	63.41	65.36	76.71	65.59	59.11
Gravel Region	52.26	54.27	54.32	56.07	57.12	63.35	57.24	54.20
Boggy Region	51.73	54.51	53.87	55.63	55.98	56.76	55.65	53.76
Vegetation Misc	64.11	66.72	66.79	68.31	70.90	71.68	70.57	66.41
Crag-stone	75.54	77.10	77.25	78.57	79.36	80.13	79.08	76.13
Road Grime	59.41	62.37	61.10	62.58	64.31	65.82	64.71	61.27
Drainage Grate	53.26	55.51	57.77	58.21	58.82	59.54	61.43	55.47
Earthen Region	63.93	66.45	66.90	67.12	67.72	69.84	67.20	64.19
Water Puddle	62.21	64.92	64.40	66.91	67.56	72.43	66.05	64.47
Misc	61.04	63.26	63.88	65.39	66.92	68.49	66.38	63.51
Concrete Region	55.34	57.53	57.17	58.45	59.71	62.86	61.78	57.91
Speed Breaker	45.87	47.15	47.46	49.44	51.37	55.35	54.24	44.58
MIoU	55.19	63.15	57.33	58.90	59.74	62.86	60.53	61.56

Table 4 Evaluation of EDS-ST on R^2S100K with different subsets of real and pseudo-labeled data

Model	Real	Pseudo	w/o EDS					w/EDS				
			Accuracy	Precision	Recall	F1-score	MIoU	Accuracy	Precision	Recall	F1-Score	MIoU
Teacher	3K	–	58.32	62.34	55.45	58.62	56.14	63.23	68.45	57.32	62.49	62.86
Student	3K	2K	59.76	61.32	58.79	60.04	59.87	64.45	69.78	58.23	63.39	63.15
Student	3K	4K	62.29	66.67	59.81	63.09	62.24	66.12	70.18	62.14	66.02	65.82
Student	3K	8K	64.87	69.75	61.12	66.37	63.50	66.95	71.23	65.45	68.23	66.03
Student	3K	16K	62.59	67.32	59.84	63.18	62.41	66.83	69.67	62.22	65.76	66.91
Student	3K	32K	62.68	68.93	57.34	64.67	**62.33**	67.43	69.78	65.21	67.43	**67.40**

The bold values indicates the best Mean IoU

4.4 Effectiveness of EDS-Based Self-Training

Following supervised learning, we used STDS and EDS to analyze the efficient training and its impact on the inference of student models. The results are summarized in Table 4, and we have several observations. Firstly, EDS significantly improves student models with an average increase of 4% MIoU. Therefore, using EDS for training segmentation models is better than not using it. Secondly, EDS can be used as a generic approach to train teacher methods efficiently. From Fig. 11, it is clear that EDS improves teacher method by 4%. Thirdly, EDS is necessary to achieve better results when pseudo labels dominate the training set such as the 16k/32k set, otherwise, the performance of the models starts declining. For instance, student models trained without EDS over 16k, and 32k pseudo labeled sets dropped by 0.8% because of redundant training samples which contribute bias towards classes with more pixels against classes with lesser ones. Whereas EDS efficiently handles data imbalance, thus it improves the performance of student models as compared to the STDS approach, as shown in Fig. 12.

In addition, student models with more pseudo labels (16K, 32K) marginally improve compared to models with lesser pseudo labels (2K/4K). With fewer pseudo labels, the model learns more informative features as variable data samples are clustered based on similar representations by EDS. However, in the case of more pseudo labels, a vast range of sequential data samples is selected from each cluster, which causes the model to start saturating instead of learning new information. On the other hand, EDS ensures the selection of distinct sampling and helps the model refine mask boundaries, ultimately benefitting dense tasks.

4.5 Comparison with Related Self-Training Methods on R^2S100K, Cityscapes, and CamVid

Here we describe a comparative analysis of existing self-training methods. As shown in Table 5, our EDS outperforms other self-training methods (Abdalla et al., 2019; Lee et al., 2013; Wang et al., 2022; Zhao et al., 2023; Zou et al., 2019, 2018) on $R^2$100K, as well as on cityscapes and CamVid. On $R^2$100K, consistency regularization achieved 53.70% mIoU i.e., considerably worse than all of the self-training methods, as the model is learning from inaccurate predictions in the first stage of training, leading to inaccurate inference on test data. Similarly, in the case of teacher fine-tuning, we observe that the model gets stuck at minima at an early stage of fine-tuning. Resultantly, the model starts overfitting instead of learning new information. Similarly, we notice that Wang et al. (2022) struggles to distinguish hazardous road regions in R^2S100K due to higher textural similarities among classes, leading to a higher misclassification rate. We first efficiently select training data samples using the EDS approach to train a teacher model with considerable accuracy and use it to produce pseudo labels of our unlabeled data. Therefore, its performance consistently improves throughout the training process. Our framework is purely generic; using our approach, a teacher model can train any student model irrespective of their architectural differences, showing its generalization capability. The performance of EDS is shown in Table 6.

4.6 Generalization to Other Student Methods

Another benefit of EDS-based self-training is that teacher and student models do not need the same architectures. Our framework is a generic pipeline that clusters data based on representations. Then, data samples are uniformly selected to ensure data balance for training a teacher model—used to generate pseudo labels which are utilized in improving the accuracy of the student model. In particular, we used DeepLabV3+ with ResNet101 as a teacher model and trained several student models including BiSeNet, PSPNet, LRASPP, LinkNet, FeedFormer, SegNeXt, and U-MixFormer with different backbone networks. These models are selected after analyzing their wide adaptation to segmentation tasks. The results in Table 7 demonstrate that EDS-based self-training can significantly improve student models irrespective of their architectures. Comparatively,

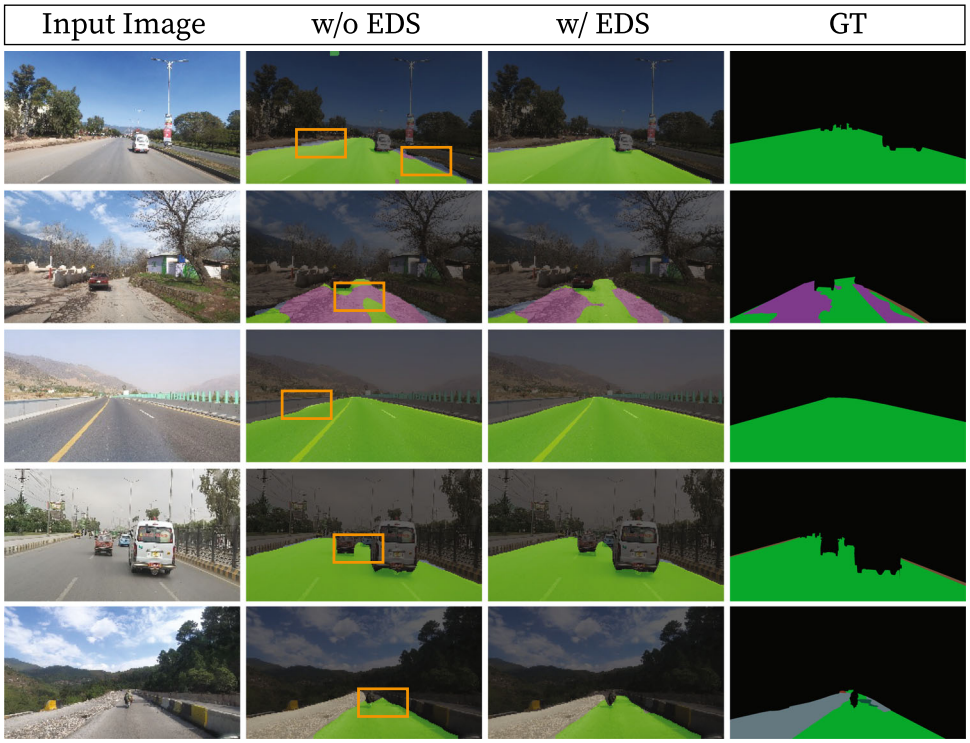

Fig. 12 Visualizing the comparison of best-performing student methods on R^2S100K. Results demonstrate that EDS-based self-training is a better approach to effectively handle class confusion in complex road scenarios. The labels are the same as in Fig. 1

Table 5 Evaluation of self-training methods on R^2S100K

Methods	Model	MIoU
Teacher Fine-tuning (Abdalla et al., 2019)	Single Model	59.21
Consistency Regularization (Zou et al., 2018)	Single Model	53.70
Model Regularizer (Zou et al., 2019)	Student + Teacher	57.64
Pseudo-labels (Lee et al., 2013)	Student + Teacher	61.05
U^2PL (Wang et al., 2022)	Student + Teacher	64.29
iMAS (Zhao et al., 2023)	Student + Teacher	66.42
EDS (Our method)	Student + Teacher	**67.40**

VThe bold values indicates the best Mean IoU

Table 6 Analyzing semi-supervised methods on R^2S100K

Methods	Venue/Year	R^2S100K		CamVid		Cityscapes	
		w/o EDS	w EDS	w/o EDS	w EDS	w/o EDS	w/EDS
Baseline (Chen et al., 2018)	CVPR 18	62.91	64.27	60.82	64.44	62.21	64.73
CRST (Zou et al., 2019)	ICCV 19	63.42	63.79	59.37	59.66	60.57	63.86
HLCon (Mittal et al., 2019)	TPAMI 19	66.38	66.91	62.13	63.71	63.94	64.18
CCT (Ouali et al., 2020)	CVPR 20	65.14	65.40	63.73	66.10	63.75	65.43
PseudoSeg (Zou et al., 2020)	ICLR 21	64.89	66.73	63.58	64.35	64.60	65.98

PSPNet with ResNet101 outperformed other segmentation networks using the EDS approach.

5 Discussion and Future Direction

The annual report of the World Economic Forum (WEF, 2019) indicates that the quality of road infrastructure of Pakistan achieved a 4 road quality index, which is similar or comparable with many high-, middle-, and low-income countries. This list includes but is not limited to India, Russia, Kuwait, Romania, Indonesia, Bulgaria, Brazil, Malta, Iceland, Hungary, Czechia, Slovakia, Ukraine, Moldova, Georgia, Jordan, and Venezuela. These statistics indicate that models trained on R2S100K can be generalized to the road

Table 7 Generalizability of student methods irrespective of different backbone network architectures on $R^2$100K

Model	Backbone	Val	Test				
		MIoU	Accuracy	Precision	Recall	F1-Score	MIoU
BiSeNet (Yu et al., 2018)	ResNet-50	62.32	63.45	68.95	56.23	62.01	63.17
BiSeNet (Yu et al., 2018) w/EDS	ResNet-50	64.40	65.33	69.75	60.82	65.17	64.93
PSPNet (Zhao et al., 2017)	ResNet-101	62.77	64.43	68.78	59.46	63.82	64.51
PSPNet (Zhao et al., 2017) w/EDS	ResNet-101	65.82	67.56	69.98	61.87	65.76	67.23
LRASPP (Howard et al., 2019)	MobileNet-v3	59.11	60.53	62.21	58.45	60.24	59.87
LRASPP (Howard et al., 2019) w/EDS	MobileNet-v3	60.56	61.45	59.57	63.12	61.23	61.38
LinkNet (Chaurasia & Culurciello, 2017)	ResNet-101	62.48	63.62	68.59	57.83	62.78	63.59
LinkNet (Chaurasia & Culurciello, 2017) w/EDS	ResNet-101	64.25	64.87	70.23	62.45	66.12	64.72
FeedFormer (Shim et al., 2023)	–	63.27	67.56	70.32	62.89	67.23	64.17
FeedFormer (Shim et al., 2023) w/EDS	–	64.53	66.92	68.54	65.78	67.13	65.97
SegNeXt (Guo et al., 2022)	–	62.98	64.53	67.24	60.32	63.89	64.52
SegNeXt (Guo et al., 2022) w/EDS	–	64.21	65.83	69.87	62.45	66.02	65.87
U-MixFormer (Yeom & von Klitzing, 2023)	–	63.52	65.22	69.64	61.12	65.19	64.03
U-MixFormer (Yeom & von Klitzing, 2023) w/EDS	–	65.98	67.85	67.52	64.23	65.85	65.91

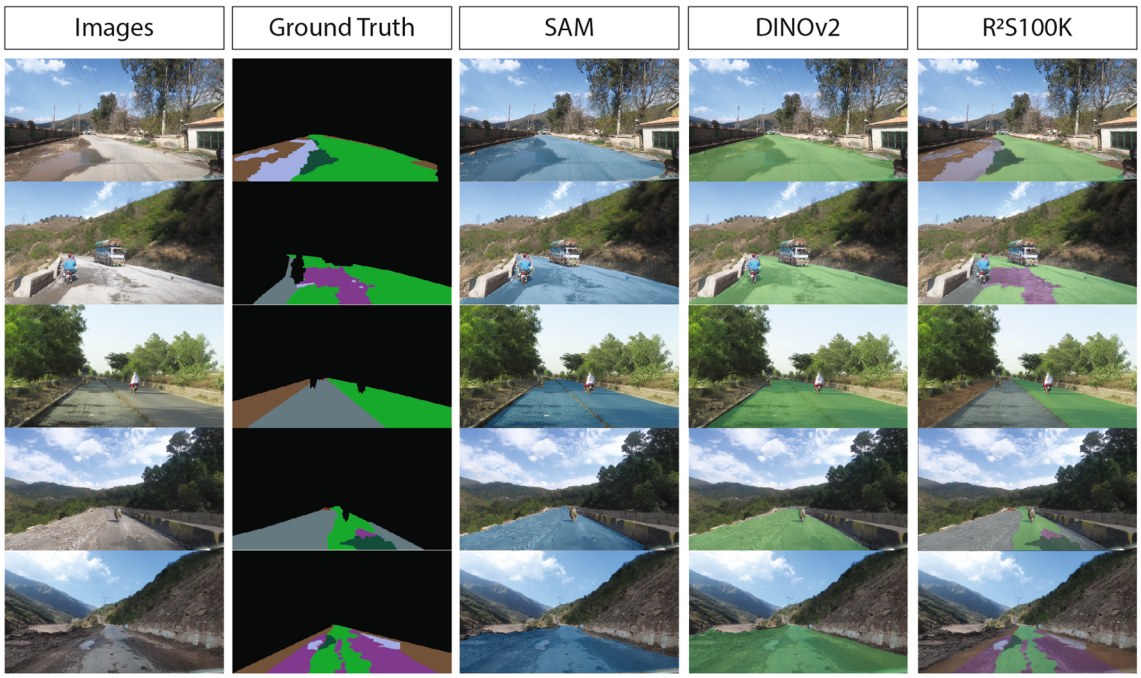

Fig. 13 Comparison between our method, Segment Anything Model and DINOv2. The colors of labels in R2S100K examples are the same as in Fig. 1

infrastructure of the aforementioned and other countries with similar road quality indexes. According to the Global Status Report on Road Safety 2023 (WHO, 2023), reporting countries collectively account for nearly 68 million km of roads, of which 4.5 million km are paved expressways; 47 million km are paved interurban roads; and 10 million km are unpaved inter-urban roads. Among these, 80% of the roads of the reporting countries do not meet a minimum 3-star rating for user safety due to non-standard road infrastructure. Consequently, 92% of deaths due to road fatalities occur in low- and middle-income countries which share similar socio-economic status.

The R^2S100K dataset provides a diverse set of road images covering challenging roadways, including hazardous road patches that are common in developing countries. This diversity enhances the utility of the dataset for training and evaluating autonomous driving perception systems. By distinguishing safe asphalt road regions from hazardous road

regions, the dataset offers finer-grained semantic segmentation labels, which can improve the accuracy of perception models in distinguishing between different types of road surfaces. Moreover, R^2S100K addresses the underrepresentation of challenging road scenarios in existing datasets, thereby improving the efficiency of autonomous driving research by providing a more comprehensive benchmark for evaluating perception models.

The R^2S100K is a first attempt at providing the labeling for semantic road region segmentation tasks. The current version of R^2S100K contains one hundred thousand images, including 15K labeled and 85K unlabeled images. It is well-known that pixel-by-pixel image labeling is costly and time-consuming; this research gap is addressed using our EDS-based self-training method. With the recent advances in computer vision, in particular with the release of foundation segmentation models, i.e., Segment Anything (Kirillov et al., 2023), DINOV2 (Oquab et al., 2023), and Internimage (Wang et al., 2023) are being adopted by the community for auto-labeling. Though these models have achieved significant performance on well-known classes, However, these models under-perform on R2S100K, and cannot segment the classes in R2S100K, as shown in Fig. 13, due to the underrepresentation of such data in SA-dataset: it only covers 0.9% of data samples from low-income countries (Kirillov et al., 2023). Therefore, we manually labeled the data utilizing our data engineers' expertise and proposed an EDS-based self-training method to efficiently utilize the unlabeled data in improving the model. In the future, we aim to integrate additional information modalities like odometry and Lidar point clouds into the dataset.

6 Conclusions

In this paper, we presented R^2S100K to perform drivable road region segmentation on unstructured roadways. We also presented a self-training framework to improve semi-supervised learning for segmentation tasks. Results demonstrate that our proposed method can be utilized to improve supervised/semi-supervised learning for semantic segmentation due to its effective class confusion handling in complex road environments. Our training framework will facilitate research in various ML applications where generating labeled data is critical. In the future, we will extend the annotations to encompass lane markings, the surrounding environment and infrastructure, and vehicles.

Acknowledgements This research was made possible by NPRP grant # [13S-0206-200273] from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

Author Contributions Muhammad Atif Butt: Conceptualization, Methodology, Writing—original draft, Writing—review & editing. Adnan Qayyum: Conceptualization, Methodology, Writing—review & editing. Hassan Ali: Conceptualization, Methodology, Writing—review & editing. Waqas Sultani: Conceptualization, Methodology, Writing—review & editing. Ala Al-Fuqaha: Conceptualization, Methodology, Funding, acquisition, Writing—review & editing. Junaid Qadir: Conceptualization, Methodology, Funding acquisition, Writing—review & editing.

Funding Open Access funding provided by the Qatar National Library.

Data Availibility R^2S100K Dataset and codes will be made publicly available on our project web page https://r2s100k.github.io/ after the publication of this manuscript.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abdalla, A., Cen, H., Wan, L., Rashid, R., Weng, H., Zhou, W., & He, Y. (2019). Fine-tuning convolutional neural network with transfer learning for semantic segmentation of ground-level oilseed rape images in a field with high weed pressure. *Computers and Electronics in Agriculture, 167*, 105091.

Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). SegNet: A deep convolutional encoder–decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(12), 2481–2495.

Brostow, G. J., Fauqueur, J., & Cipolla, R. (2009). Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters, 30*(2), 88–97.

Budvytis, I., Sauer, P., Roddick, T., Breen, K. & Cipolla, R. (2017). Large scale labelled video data augmentation for semantic segmentation in driving scenarios. In *Proceedings of the IEEE international conference on computer vision workshops* (pp. 230–237).

Butt, M. A., & Riaz, F. (2022). CARL-D: A vision benchmark suite and large scale dataset for vehicle detection and scene segmentation. *Signal Processing: Image Communication, 104*, 116667.

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G. & Beijbom, O. (2020). nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11621–11631).

Chaurasia, A. & Culurciello, E. (2017). LinkNet: Exploiting encoder representations for efficient semantic segmentation. In *2017 IEEE*

visual communications and image processing (VCIP) (pp. 1–4). IEEE.

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. (2014). Semantic image segmentation with deep convolutional nets and fully connected CRFs. arXiv preprint arXiv:1412.7062

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2017a). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*(4), 834–848.

Chen, L. C., Papandreou, G., Schroff, F., & Adam, H. (2017b). Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587

Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder–decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)* (pp 801–818).

Chen, X., Yuan, Y., Zeng, G., & Wang, J. (2021). Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp 2613–2622).

Cheng, B., Schwing, A., & Kirillov, A. (2021). Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems, 34*, 17864–17875.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3213–3223).

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). IEEE.

Dollár, P., Wojek, C., Schiele, B., & Perona, P. (2009). Pedestrian detection: A benchmark. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 304–311). IEEE.

Gaigbe-Togbe, V., Bassarsky, L., Gu, D., Spoorenberg, T., & Zeifman, L. (2022). *United nations world population prospects*. UNCTAD.

Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? The kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition* (pp. 3354–3361). IEEE.

Geyer, J., Kassahun, Y., Mahmudi, M., Ricou, X., Durgesh, R., Chung, A.S., Hauswald, L., Pham, V.H., Mühlegg, M., Dorn, S. & Fernandez, T. (2020). A2d2: Audi autonomous driving dataset. arXiv preprint arXiv:2004.06320

Guo, M. H., Lu, C. Z., Hou, Q., Liu, Z., Cheng, M. M., & Hu, S. M. (2022). SegNeXt: Rethinking convolutional attention design for semantic segmentation. *Advances in Neural Information Processing Systems, 35*, 1140–1156.

He, T., Shen, C., Tian, Z., Gong, D., Sun, C., & Yan, Y. (2019). Knowledge adaptation for efficient semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp 578–587).

Howard, A., Sandler, M., Chu, G., Chen, L. C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., & Le, Q. V. (2019). Searching for mobileNetV3. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp 1314–1324).

Huang, X., Cheng, X., Geng, Q., Cao, B., Zhou, D., Wang, P., Lin, Y., & Yang, R. (2019). The apolloscape open dataset for autonomous driving and its application. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42*(10), 2702–2719.

Hung, W. C., Tsai, Y. H., Liou, Y. T., Lin, Y. Y., & Yang, M. H. (2018). Adversarial learning for semi-supervised semantic segmentation. arXiv preprint arXiv:1802.07934

Jiang, P., Osteen, P., Wigness, M., & Saripalli, S. (2021). RELLIS-3D dataset: Data, benchmarks and analysis. In *2021 IEEE international conference on robotics and automation (ICRA)* (pp. 1110–1116). IEEE.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W. Y., & Dollár, P. (2023). Segment anything. arXiv preprint arXiv:2304.02643

Lee, D. H. (2013). Pseudo-Label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML* (p. 896).

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *European conference on computer vision* (pp. 740–755). Springer.

Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2117–2125).

Liu, W., Rabinovich, A., & Berg, A. C. (2015). ParseNet: Looking wider to see better. arXiv preprint arXiv:1506.04579

Liu, Y., Shu, C., & Wang, J. (2020). Structured knowledge distillation for dense prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*, 7035.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431–3440).

Luc, P., Neverova, N., Couprie, C., Verbeek, J., & LeCun, Y. (2017). Predicting deeper into the future of semantic segmentation. In *Proceedings of the IEEE international conference on computer vision* (pp 648–657).

Mittal, S., Tatarchenko, M., & Brox, T. (2019). Semi-supervised semantic segmentation with high- and low-level consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(4), 1369–1379.

Mustikovela, S. K., Yang, M. Y., & Rother, C. (2016). Can ground truth label propagation from video help semantic segmentation? In *European conference on computer vision* (pp. 804–820). Springer.

Neuhold, G., Ollmann, T., Rota Bulo, S., & Kontschieder, P. (2017) The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision* (pp. 4990–4999).

Noh, H., Hong, S., & Han, B. (2015). Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision* (pp 1520–1528).

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A. & Assran, M. (2023). Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193

Ouali, Y., Hudelot, C., & Tami, M. (2020). Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp 12674–12684).

Rasib, M., Butt, M. A., Riaz, F., Sulaiman, A., & Akram, M. (2021). Pixel level segmentation based drivable road region detection and steering angle estimation method for autonomous driving on unstructured roads. *IEEE Access, 9*, 167855–167867.

Rezaei, M., Yang, H., & Meinel, C. (2020). Recurrent generative adversarial network for learning imbalanced medical image semantic segmentation. *Multimedia Tools and Applications, 79*(21), 15329–15348.

Romera, E., Alvarez, J. M., Bergasa, L. M., & Arroyo, R. (2017). ERFNet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems, 19*(1), 263–272.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241). Springer.

Schwab, K. (2019). *Global competitiveness index 4.0 2019 edition*. World Economic Forum.

Shim, J. H., Yu, H., Kong, K., & Kang, S. J. (2023). FeedFormer: Revisiting transformer decoder for efficient semantic segmentation. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 2263–2271).

Siam, M., Gamal, M., Abdel-Razek, M., Yogamani, S., Jagersand, M., & Zhang, H. (2018). A comparative study of real-time semantic segmentation for autonomous driving. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 587–597).

Souly, N., Spampinato, C., & Shah, M. (2017). Semi and weakly supervised semantic segmentation using generative adversarial network. arXiv preprint arXiv:1703.09695

Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., & Vasudevan, V. (2020). Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2446–2454).

SuperAnnotate AI Inc. (2024). *AI data platform for LLM, CV, and NLP*. https://www.superannotate.com

Varma, G., Subramanian, A., Namboodiri, A., Chandraker, M., & Jawahar, C. V. (2019). IDD: A dataset for exploring problems of autonomous navigation in unconstrained environments. In *2019 IEEE winter conference on applications of computer vision (WACV)* (pp. 1743–1751). IEEE.

Viswanath, K., Singh, K., Jiang, P., Sujit, P. B., & Saripalli, S. (2021). OFFSEG: A semantic segmentation framework for off-road driving. In *2021 IEEE 17th international conference on automation science and engineering (CASE)* (pp. 354–359). IEEE.

Wang, W., Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., Hu, X., Lu, T., Lu, L., Li, H., & Wang, X. (2023). InternImage: Exploring large-scale vision foundation models with deformable convolutions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14408–14419).

Wang, Y., Wang, H., Shen, Y., Fei, J., Li, W., Jin, G., Wu, L., Zhao, R. & Le, X. (2022). Semi-supervised semantic segmentation using unreliable pseudo-labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4248–4257).

WEF. (2019). *Roads quality by country, around the world*. TheGlobalEconomy.com. https://www.theglobaleconomy.com/rankings/roads_quality/

Wei, Y., Xiao, H., Shi, H., Jie, Z., Feng, J., & Huang, T. S. (2018) Revisiting dilated convolution: A simple approach for weakly- and semi-supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7268–7277).

WHO. (2019). *World health statistics overview 2019: Monitoring health for the SDGs, sustainable development goals*. Tech. rep: World Health Organization.

WHO. (2020). *World health statistics 2020*. Tech. rep: World Health Organization.

WHO. (2023). *Global status report on road safety 2023*. https://www.who.int/teams/social-determinants-of-health/safety-and-mobility/global-status-report-on-road-safety-2023

Wigness, M., Eum, S., Rogers, J. G., Han, D., & Kwon, H. (2019) A RUGD dataset for autonomous navigation and visual perception in unstructured outdoor environments. In *2019 IEEE/RSJ international conference on intelligent robots and systems (IROS)* (pp. 5000–5007). IEEE.

Xiao, P., Shao, Z., Hao, S., Zhang, Z., Chai, X., Jiao, J., Li, Z., Wu, J., Sun, K., Jiang, K., & Wang, Y. (2021). PandaSet: Advanced sensor suite dataset for autonomous driving. In *2021 IEEE international intelligent transportation systems conference (ITSC)* (pp. 3095–3101). IEEE.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems, 34*, 12077–12090.

Xie, J., Shuai, B., Hu, J. F., Lin, J., & Zheng, W. S. (2018). Improving fast segmentation with teacher–student learning. arXiv preprint arXiv:1810.08476

Yeom, S. K., & von Klitzing, J. (2023) U-MixFormer: UNet-like transformer with mix-attention for efficient semantic segmentation. arXiv preprint arXiv:2312.06272

Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., & Sang, N. (2018). BiSeNet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 325–341).

Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., & Darrell, T. (2020). BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2636–2645).

Yu, L., Liu, X., & Van de Weijer, J. (2022). Self-training for class-incremental semantic segmentation. *IEEE Transactions on Neural Networks and Learning Systems, 34*, 9116.

Zendel, O., Honauer, K., Murschitz, M., Steininger, D., & Dominguez, G. F. (2018). Wilddash-creating hazard-aware benchmarks. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 402–416).

Zhang, S., Benenson, R., & Schiele, B. (2017). CityPersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3213–3221).

Zhang, X., Du, B., Wu, Z., & Wan, T. (2022). LAANet: Lightweight attention-guided asymmetric network for real-time semantic segmentation. *Neural Computing and Applications, 34*, 1–15.

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2881–2890).

Zhao, H., Qi, X., Shen, X., Shi, J., & Jia, J. (2018). ICNet for real-time semantic segmentation on high-resolution images. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 405–420).

Zhao, Z., Long, S., Pi, J., Wang, J., & Zhou, L. (2023). Instance-specific and model-adaptive supervision for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 23705–23714).

Zhu, Y., Sapra, K., Reda, F. A., Shih, K. J., Newsam, S., Tao, A., & Catanzaro, B. (2019). Improving semantic segmentation via video propagation and label relaxation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8856–8865).

Zou, Y., Yu, Z., Kumar, B. V. K., & Wang, J. (2018). Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)* (pp 289–305).

Zou, Y., Yu, Z., Liu, X., Kumar, B. V. K., & Wang, J. (2019). Confidence regularized self-training. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 5982–5991).

Zou, Y., Zhang, Z., Zhang, H., Li, C. L., Bian, X., Huang, J. B., & Pfister, T. (2020). PseudoSeg: Designing pseudo labels for semantic segmentation. arXiv preprint arXiv:2010.09713