PalmCity: An Emerging Benchmark Dataset for Semantic Segmentation of Panoramic Street View Images in Under-Represented Developing Countries

Muzaffer Can Iban[1(✉)], Onur Can Bayrak[2], Serkan Kartal[3], Dogu Ilmak[4], and Dursun Zafer Seker[5]

[1] Department of Geomatics Engineering, Mersin University, 33343 Mersin, Turkey
caniban@mersin.edu.tr
[2] Department of Geomatics Engineering, Yildiz Technical University, Istanbul, Turkey
onurcb@yildiz.edu.tr
[3] Department of Computer Engineering, Cukurova University, Adana, Turkey
skartal@cu.edu.tr
[4] Department of Remote Sensing and GIS, Mersin University, Mersin, Turkey
[5] Department of Geomatics Engineering, Istanbul Technical University, Istanbul, Turkey
seker@itu.edu.tr

Abstract. Street View Imagery (SVI) offers detailed, street-level data for urban analysis, enabling the study of green spaces, sky, buildings, and other urban elements through semantic segmentation. Techniques like Green/Sky/Building View Indexes link urban morphology, climate, socio-economic factors, and public health. However, pre-trained models such as Cityscapes and ADE20K, designed for cities in developed countries, often fail to represent the diverse architectural and land-use patterns of developing countries like Türkiye, resulting in poor segmentation performance. To address this, the PalmCity project introduces a tailored benchmark dataset for Türkiye's unique urban characteristics. Using 360-degree action cameras, PalmCity will collect at least 5,000 panoramic SVI images from Mersin City, chosen for its representative urban typologies. The dataset aims to improve SVI semantic segmentation and support urban studies in under-represented regions. PalmCity is going to evaluate the state-of-the-art deep learning models, including FCN, PSPNet, DeepLabV3 and Transformer-based models, using ResNet as the primary backbone. Models trained on PalmCity are going to be compared to Cityscapes-trained models to assess segmentation performance. Preliminary results show that Cityscapes weights perform well for general classes like sky, road, building and trees but struggle with urban objects, vehicles, and panoramic distortions in PalmCity images, underscoring the need for dataset-specific training.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
O. Gervasi et al. (Eds.): ICCSA 2025 Workshops, LNCS 15899, pp. 165–175, 2026.
https://doi.org/10.1007/978-3-031-97663-6_14

Keywords: Street View Imagery · Benchmark Dataset · Semantic Segmentation · Deep Learning

1 Introduction

1.1 Street View Imagery

In recent years, Street View Imagery (SVI) has become a pivotal tool in urban studies, offering street-level, panoramic views of urban environments [1]. The rise of services like Google Street View (GSV), advancements in deep learning and computer vision, and increased computational power have driven the widespread adoption of SVI in academic research [2]. Unlike aerial or satellite imagery, SVI provides a horizontal, human-eye perspective, enabling detailed, human-centric analysis of urban features. This has proven invaluable in urban planning, environmental assessment, traffic optimisation, and city management, supporting sustainable urban development.

Google is the largest global SVI provider, with GSV covering over 90 countries. GSV images, captured by vehicle-mounted multi-directional cameras, offer panoramic views. Researchers can download GSV images in bulk via the Google Maps API, though limitations exist. For instance, GSV is unavailable in China, where Tencent Street View and Baidu Total View dominate [2]. Volunteer-based platforms like Mapillary and KartaView complement GSV by hosting user-contributed images, particularly useful for areas inaccessible to vehicles, such as pavements and pedestrian paths. These platforms are especially beneficial for studying developing or remote regions [3].

However, both commercial and volunteer-based SVI providers have limitations. The Google Maps API restricts access to historical images and often provides lower-resolution downloads. GSV update frequencies vary, with some areas having outdated or seasonally inconsistent images, and weather conditions further affect data quality [4]. Volunteer-based platforms, while valuable, often rely on non-panoramic dashcam images, offering less comprehensive perspectives [3]. Despite these challenges, SVI remains a critical resource for urban research, bridging gaps in traditional data collection methods.

1.2 Information Extraction from SVI

SVI data is widely used to derive key indicators like the Green View Index (GVI) and Sky View Factor (SVF), essential for assessing urban environmental conditions. GVI measures vegetation proportion in SVI images, reflecting greenery from a human perspective, while SVF quantifies visible sky, influencing urban microclimates. Together, they help monitor urban heat islands and evaluate thermal comfort. A 2022 review of 135 studies found that 62% used deep learning, 23% relied on unsupervised RGB/HSV classification, and 13% employed manual digitisation [4]. Deep learning techniques rely on semantic segmentation, which groups pixels by extracted features, to predict their position and class. Widely used in autonomous driving, medical imaging, and remote sensing, it enables

precise object identification. Popular architectures for semantic segmentation of SVI include Fully Convolutional Networks (FCN), Pyramid Scene Parsing Networks (PSPNet), SegNet, DeepLabV3, and DeepLabV3+. These semantic segmentation techniques group pixels by object, effectively extracting GVI and SVF from SVI data.

1.3 Limitations of Current Benchmark Datasets

The majority of SVI studies rely on pre-trained deep learning models that have been trained on benchmark datasets. Among these, Cityscapes and ADE20K are the most commonly used. However, these datasets have notable limitations that affect their applicability to urban research, particularly in developing countries.

ADE20K. Developed by MIT's Computer Science and Artificial Intelligence Laboratory, ADE20K contains scenes representing various environments but is not specifically focused on street views [5]. It includes 150 semantic classes, many of which pertain to indoor objects, making it less suitable for street-level analysis.

Cityscapes. Created by TU Darmstadt, the Max Planck Institute for Informatics and Daimler, Cityscapes is designed for autonomous vehicle navigation and includes 30 object classes relevant to driving scenarios. It comprises 5,000 finely annotated and 20,000 coarsely annotated images captured from vehicle-mounted cameras in 50 cities in Germany and neighbouring countries [6]. While Cityscapes is the first benchmark dataset dedicated to street views, it has several limitations. The images in the data set are captured using a 2-megapixel camera, which may not provide sufficient resolution for detailed segmentation tasks. Furthermore, the fixed perspective of the camera limits the generalisation ability of models trained on this dataset, particularly for urban research requiring panoramic views. Additionally, Cityscapes does not include semantic classes for water bodies, staircases, overpasses, or certain types of vegetation, which are critical for comprehensive urban analysis.

Other Datasets. Such as CamVid, KITTI, Mapillary Vistas, DarkZurich, and BDD100K, are also used for the semantic segmentation of street scenes. However, these data sets are primarily focused on driving and traffic scenarios, limiting their applicability to pedestrian-focused urban research. Moreover, the variability in image quality due to the use of different devices and the imbalance in class distribution further constrain their utility [7].

1.4 PalmCity Project

Benchmark datasets are often pre-trained on models that include weights for the classes present in the dataset. Although this allows users to predict semantic

classes in their own images without additional annotation, the generalisation of these models to different urban contexts, particularly in developing countries, remains a challenge [8]. For example, Cityscapes and similar datasets predominantly represent European and North American urban patterns, which may not adequately capture the architectural and land use diversity found in cities in developing countries [9]. This highlights the need for region-specific datasets that reflect the unique urban characteristics of these areas.

Urban areas in developing countries, particularly Türkiye, exhibit common patterns in land use and architectural diversity. This study identifies six distinct typologies present in Turkish cities: market areas, planned residences, planned apartments, informal settlements, industrial urban areas, and coastal areas [10]. Mersin has been selected as the primary case study, as it contains all six typologies and serves as a suitable representative for other Turkish cities. The SVI data collected from Mersin will be used to develop transferable and generalisable models applicable to other urban areas in developing countries. Additionally, SVI scenes from other cities may be integrated into the dataset through voluntary contributions or related projects, enhancing research on semantic segmentation, object detection, and scene classification in urban environments.

Each typology reflects varying densities of urban objects, materials, human and vehicle behaviours, and distributions of green spaces and sky visibility. However, current datasets used for autonomous vehicle navigation lack consideration of these architectural and land-use typologies, primarily focusing on pedestrian and vehicle detection. Moreover, large-scale datasets are often deemed necessary for generalisation in real-world applications, yet their preparation and processing present significant challenges. Recent studies suggest that domain adaptation and few-shot learning offer effective alternatives, reducing dataset size while maintaining model performance. Few-shot learning enables deep learning models to generalise from limited training examples, while domain adaptation allows models trained on one dataset to perform effectively in different environments [11,12].

The PalmCity project aims to create an urban dataset that addresses the limitations of existing semantic segmentation datasets by capturing a diverse range of urban patterns rather than a monotonous built environment. It ensures the structured collection and annotation of urban scenes to reflect typological variations. Additionally, it provides a more efficient and adaptable dataset, reducing the need for extensive manual annotation while maintaining high model performance.

The PalmCity dataset, derived from Mersin, is designed to offer a richer variety of object classes and urban features compared to existing datasets, particularly in developing countries. Key differences include: (i) PalmCity includes streets with inconsistent lane markings, variable pavement widths, and unpaved surfaces. (ii) Developing countries exhibit greater visual variability in vehicles due to wear, modifications, and informal traffic behaviours. Pedestrians frequently cross outside designated areas, and motorcycles are widely used. (iii) PalmCity features dense advertising billboards, informal vendors, and horse-drawn carts, reflecting local urban life.

By capturing these elements, the PalmCity dataset enhances autonomous vehicle research, urban transformation studies, and security assessments in complex urban environments. Unlike traditional datasets that prioritise large-scale image collection, PalmCity focuses on strategic scene selection, ensuring greater typological representation and efficiency in model training.

2 Materials and Methods

2.1 Field Work

The fieldwork for PalmCity project was meticulously planned and executed to ensure the collection of high-quality 360° panoramic SVI across selected neighbourhoods in Mersin, Türkiye. The primary objective was to capture a diverse range of urban landscapes, reflecting variations in land use and architectural typologies, as indicated in the previous section. Four districts within Mersin city centre—Yenişehir, Akdeniz, Toroslar, and Mezitli—were prioritised, with specific neighbourhoods such as Hürriyet, Çiftlikköy, İnönü, Çankaya, Sağlık, and Viranşehir identified as key locations for data collection. At least 5000 panoramic SVI images will be obtained.

The fieldwork is being conducted using GoPro Max 360 action cameras, chosen for their dual-lens design and ability to capture 16.6-megapixel 360° panoramic images. The camera is mounted on a selfie stick attached to a waist harness, mimicking human eye-level perception (see Fig. 1). This configuration allows for stable and consistent image capture while moving through urban environments. To ensure accurate geotagging, the camera is synchronised with smartphones to record location data in real time. Images are stored on Micro SD cards and later transferred to computers via Type-C cables or Wi-Fi.

Field activities are carried out during clear weather conditions, avoiding rainfall or overcast skies, to maintain image quality. Shooting takes place during midday when sunlight was optimal, ensuring minimal shadows and consistent lighting. The camera is positioned at approximately 2.2 m above ground level. A team of assistants accompanies the operator to ensure safety, particularly when capturing images from the centre of roads. A distance of 15-âĂŞ20 m is maintained between consecutive shots, with adjustments made at intersections or curved streets to ensure comprehensive coverage of the urban landscape.

2.2 Annotation Protocol

The annotation protocol for this project was designed to ensure the accurate and consistent semantic annotation of at least 1,000 selected SVI images, reflecting the complexity and diversity of urban environments in Mersin, Türkiye. The annotation process is being conducted using the Supervisely web platform, which supports simultaneous collaboration among multiple users and allows for the annotation of up to 5,000 images. It also enables the export of annotated mask files in multiple formats, including COCO, Pascal VOC, and Cityscapes, ensuring compatibility with various machine learning frameworks.

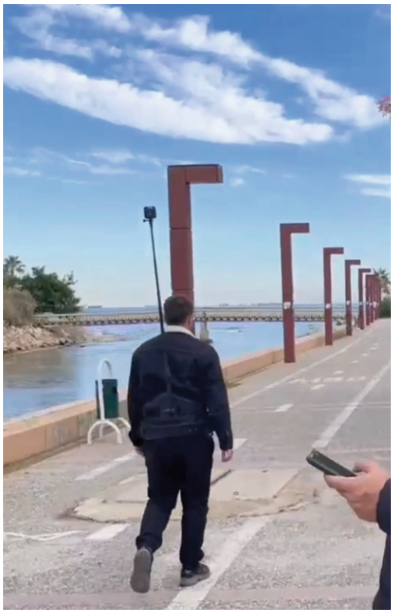

Fig. 1. Image collection with a 360-degree action camera.

The annotation process involves a team of researchers and project assistants (8 annotators, 2 admins), who meticulously annotate each image using a predefined protocol. The annotation begins with background objects and progressed to foreground elements to avoid redundant annotations. Each image requires approximately 2 h and 30 min for annotation, followed by an additional 30 min for quality control. To ensure accuracy, a subset of 20 randomly selected images will be re-annotated by different team members, and the results were compared using metrics such as the Intersection over Union (IoU) for boundary precision and F1 score for label (class) consistency.

Two class hierarchies are being established: a general first-level hierarchy (e.g., buildings, sky, vegetation) and a detailed second-level hierarchy for segmentation model development. The first-level classes will be used for factor calculations (GVI and SVF), while the second level will support advanced segmentation tasks. Table 1 provides a detailed list of second-level semantic classes. A *void* category was included to account for unlabellable or negligible objects, ensuring that segmentation models were not trained on irrelevant pixels. Standard RGB colour codes were assigned to each class to maintain visual consistency and facilitate comparison with datasets like Cityscapes.

Post-annotation, statistical analyses will be conducted to evaluate the dataset's quality and diversity. These included volumetric and density calculations to determine the spatial distribution of semantic classes, object count analyses to assess diversity, and the calculation of void pixel percentages.

Table 1. The semantic object classes considered in the labelling studies, along with their descriptions.

Semantic Class	Description
Road	The surface on which vehicles are driven. Includes all lanes, directions, and street types. Areas separated by texture changes and markings (e.g., bicycle lanes) are also roads. Does not include kerbstones.
Sidewalk	Surfaces separated by kerbstones, posts, or markings for pedestrians to walk on.
Parking Lot	Areas marked with distinct patterns and designated for vehicle parking, different from roads.
Parking Barrier	The products applied in all places to prevent vehicle parking violations.
Soil	Bare soil cover found in empty plots or under plants at the base of pavements.
Pedestrian	Only people walking or standing still in the scene. Includes people carrying handcarts and pushchairs.
Driver	People on or inside vehicles in the scene (drivers, visible passengers, etc.)
Car	Cars, jeeps, SUVs, small vans, open and closed commercial vehicles, and caravans.
Truck	Heavy-duty trucks and lorries carrying loads and containers (includes loads and containers).
Bus	All buses, minibuses, and midibuses carrying more than 9 passengers (includes public transport vehicles).
Motorcycle	Motorcycles, mopeds, and scooters. Excluding riders.
Bicycle	Bicycles and tricycles powered by muscle. Excluding riders.
Traffic Light	All illuminated traffic signs and lights, excluding poles.
Traffic Sign	All direction, parking, and traffic rule signs. Front faces of signs with text and symbols will be annotated. Back faces are void.
Pole	Vertically and rarely horizontally positioned poles carrying traffic lights, signs, and urban lighting.
Garbage Box	Street waste containers and small rubbish bins.
Sitting Bench	Benches and seats mounted on streets for resting purposes.
Infrastructure Cover	Covers on the ground providing access to electricity, water, natural gas, sewage, and clean water infrastructure.
Infrastructure Box	Electricity, natural gas, and telephone infrastructure boxes on the ground.
Building	Includes all windows, glass surfaces, doors, railings, awnings and umbrellas, signs, external air conditioning units, satellite dishes, roofs, terraces, balconies, clotheslines, and all objects on them that make up a building.
Wall	Includes retaining walls and protective walls built around buildings. Does not include walls forming the exterior surface of the building.
Fence	All fences made of metal wire or wood.
Stairs	Stairs visible in the scene, located on streets and building entrances.
Railing	All types of railings and roadside crash barriers.
Overpass	Overpasses designated for pedestrian use over roads. Does not include highway overpasses.
Water Surfaces	Sea surfaces found in coastal panoramas. Includes boats.
Sky	Open sky and clouds. Includes electrical wires in front of the sky and birds that may be present in the view.
Tree	Trees, plant-based hedges, and all types of vertical vegetation. Does not include pots on balconies and windows. Includes vegetation covering a significant portion of a wall or building surface.
Grass	Herbaceous and horizontal vegetation found in undeveloped plots.
Pruned Tree	Trees specifically pruned or deciduous.
Operator	The head and body of the operator carrying the camera appear at the bottom of the image due to cylindrical projection. These pixels are not annotated as pedestrians.
Void	Objects that will not be in the scene the next day/hour. Suitcases and bags, rubbish bags, chairs, street animals, wheelchairs. Indistinguishable scattered objects in the image. Mountains, hills, small objects that are very difficult to annotate, backs of signs.

2.3 Training Semantic Segmentation Models on PalmCity Data Set

After the labelling process is completed, the images will be divided into training, testing, and validation subsets at a ratio of 0.60:0.30:0.10. This division will not be random but will ensure an even distribution based on image locations, land use/architectural typologies, and capture times. The goal is to obtain at least 600 training, 300 testing, and 100 validation samples, though these ratios are flexible and can be optimised to meet the project's objectives. A key step in developing a semantic segmentation system is selecting the ourperforming model. To achieve this, the performance of various model architectures and backbone variations will be thoroughly evaluated. The architecture determines the neural network's structure and design, enabling the system to identify and classify image segments. Encoder-decoder structures are commonly used, with convolutional layers extracting features and decoders merging these features to make predictions.

A total of nine deep neural network architectures is going to be tested on the PalmCity dataset, including established models like FCN, PSPNet, DeepLabV3, and DeepLabV3+, as well as novel architectures such as DDRNet, PIDNet, Mask2Former, K-Net, and SegFormer. The project will also remain open to incorporating emerging architectures. Training is going to primarily use ResNet as the backbone, with controlled experiments conducted using alternative backbones for comparison. The models is going to be trained using leading deep learning libraries like PyTorch and TensorFlow, with the training set teaching the models to distinguish segmentation classes. The validation set will prevent overfitting, while the test set will evaluate performance using metrics like Pixel Accuracy and Intersection over Union (IoU). Trained model weights is going to be shared for top-performing models. Additionally, models trained on the PalmCity dataset will be compared with pre-trained Cityscapes models to assess segmentation performance, with mIoU values used as a benchmark. This comparison will help determine the effectiveness of the PalmCity dataset relative to Cityscapes.

3 Preliminary Results

As of March 2025, the PalmCity project, which began in November 2024, has completed 50% of the fieldwork for image acquisition and 30% of the annotation tasks. The exported annotation files have been tested for suitability in semantic segmentation tasks.

Using weights pre-trained on the Cityscapes benchmark dataset, semantic segmentation predictions were generated for 10 images from the PalmCity dataset, and their performance was assessed using the IoU metric. Segmentation was performed for 17 semantic classes, compatible with both the Cityscapes and PalmCity datasets, using 13 different deep learning models. An example of segmented images and their results are shown in Fig. 2, while the IoU values for specific semantic classes and the mean IoU (mIoU) values across all classes are presented in Fig. 3.

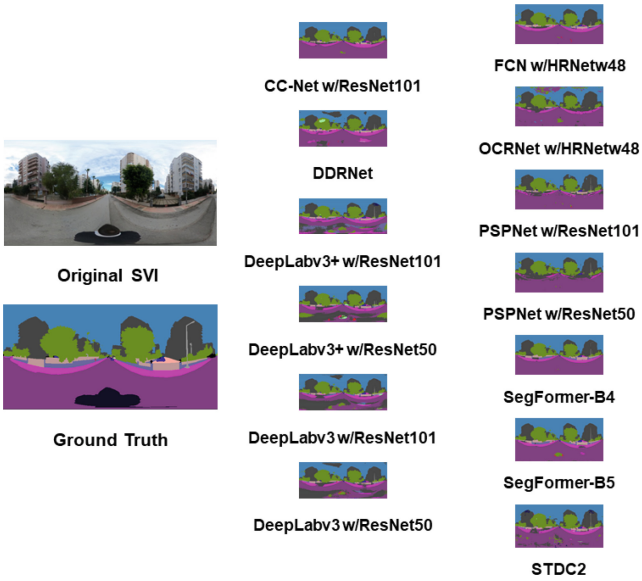

Fig. 2. Semantic segmentation predictions on PalmCity images using deep learning models pre-trained on the Cityscapes benchmark dataset.

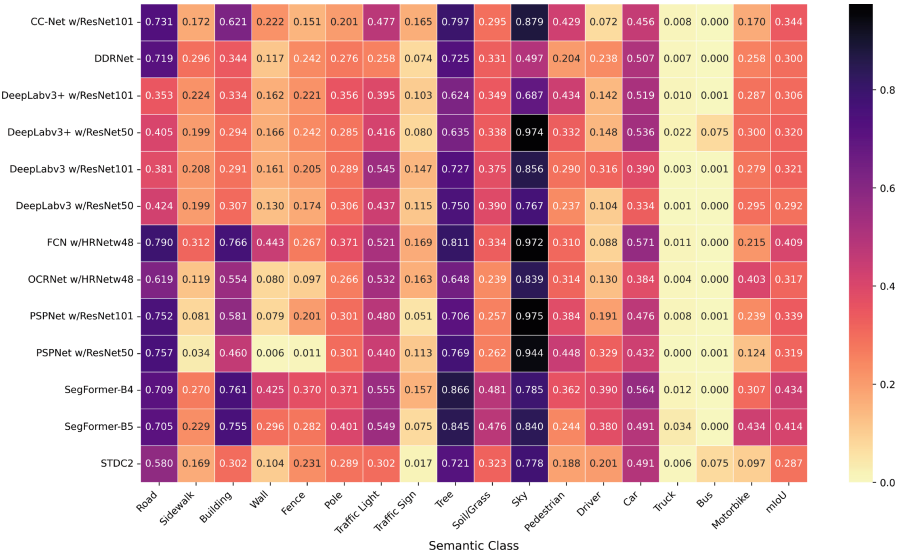

Fig. 3. Class-based semantic segmentation results.

As evident from the performance results in Fig. 3, the weights from the Cityscapes benchmark dataset provide acceptable performance for generalised semantic classes such as sky, road, building and trees in the PalmCity dataset

images. However, for classes like various vehicle types, urban objects, traffic signs, pavements, and parking areas, the generalisation capability of the Cityscapes dataset falls short when applied to PalmCity images. Additionally, the panoramic nature of PalmCity SVI images, which causes certain objects to appear distorted due to cylindrical projection, further limits the generalisation potential of the Cityscapes dataset. Therefore, the need for the PalmCity dataset is justified by several factors: the diversity of urban objects in developing countries, the presence of panoramic scenes, and the fact that the action cameras used for PalmCity offer eight times higher resolution detail compared to those used for Cityscapes. Moreover, the semantic segmentation performance of classes unique to the PalmCity dataset, which are not included in Cityscapes, can only be evaluated after training and testing models specifically on the PalmCity dataset.

4 Impact

A competition will be organized to introduce and promote the PalmCity dataset in the data science community, with the goal of increasing its use in model development, practical applications, and academic connections. The dataset will be split into training, validation (with images and annotation files) and test sets (with raw images). Participants will use the training set to develop models, and their predictions will be compared to hidden test segmentation annotations on a platform like Kaggle. Results will be evaluated on a leader board, with a secondary leader board kept secret to prevent participants from manipulating outcomes. Detailed information about the competition, rules, and data sharing is going to be available on the project website.

Additionally, several events will be organized: a training session at Mersin University to discuss the dataset's contributions, a meeting at Istanbul Technical University to promote open access with stakeholders (Mapillary, WikiMedia, OpenStreetMap etc.), an online meeting with international researchers, a workshop at Mersin University to integrate PalmCity with datasets from other disciplines, and presentations at both national and international conferences. These activities will continue from the 18th month until the project's conclusion.

Acknowledgments. This study was supported by the Scientific and Technological Research Council of Türkiye (TÜBİTAK) 3501 - Career Development Program (CAREER) under the Grant Number 124Y224. We would like to thank TÜBİTAK for their support in the project.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Biljecki, F., Ito, K.: Street view imagery in urban analytics and GIS: a review. Landsc. Urban Plan. **215**, 104217 (2021)

2. Zhang, F., Wu, L., Zhu, D., Liu, Y.: Social sensing from street-level imagery: a case study in learning spatio-temporal urban mobility patterns. ISPRS J. Photogramm. Remote. Sens. **153**, 48–58 (2019)
3. Juhász, L., Hochmair, H.H.: User contribution patterns and completeness evaluation of Mapillary, a crowdsourced street level photo service. Trans. GIS **20**(6), 925–947 (2016)
4. Lu, Y., Ferranti, E., Chapman, L., Pfrang, C.: Assessing urban greenery by harvesting street view data: a review. Urban Forestry Urban Green. **83**(6), 127917 (2023)
5. Zhou, B., et al.: Semantic understanding of scenes through the ADE20K dataset. Int. J. Comput. Vision **127**, 302–321 (2019)
6. Cordts, M., et al.: The Cityscapes dataset for semantic urban scene understanding. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 3213–3223. IEEE (2016)
7. Ülkü, I., Akagündüz, E.: A survey on deep learning-based architectures for semantic segmentation on 2D images. Appl. Artif. Intell. **36**(1), 2032924 (2022)
8. Varma, G., Subramanian, A., Namboodiri, A., Chandraker, M., Jawahar, C. V.: IDD: a dataset for exploring problems of autonomous navigation in unconstrained environments. In: IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa Village, HI, USA, pp. 1743–1751. IEEE (2019)
9. Baheti, B., Innani, S., Gajre, S., Talbar, S.: Semantic scene segmentation in unstructured environment with modified DeepLabV3+. Pattern Recogn. Lett. **138**, 223–229 (2020)
10. Çalışkan, O., Temizel, N.P., Akay, M., Mashhoodi, B.: Typological diversity and morphological continuity in the modern residential fabric: the case of Ankara. Turkey. Habitat Int. **142**, 102950 (2023)
11. Zhang, J., Ma, C., Yang, K., Roitberg, A., Peng, K., Stiefelhagen, R.: Transfer beyond the field of view: dense panoramic semantic segmentation via unsupervised domain adaptation. IEEE Trans. Intell. Transp. Syst. **23**(7), 9478–9491 (2021)
12. Ankareddy, R., Delhibabu, R.: Dense segmentation techniques using deep learning for urban scene parsing: a review. IEEE Access **13**, 34496–34517 (2025)