



Research paper

On real-time semantic segmentation with comprehensive off-road datasets for enhanced terrain classification

Semih Beycimen^a, Ali Art^c, Mehmet Unal^b, Dmitry Ignatyev^b, Argyrios Zolotas^b^a Istanbul Technical University, Faculty of Electrical and Electronics, Control and Automation Engineering, Istanbul, 34485, Türkiye^b Cranfield University, Centre for Assured and Connected Autonomy, Faculty of Engineering and Applied Sciences, Cranfield, MK43 0AL, United Kingdom^c Karabuk University, Centre for Electrical and Electronics Engineering, Karabuk, 78050, Türkiye

ARTICLE INFO

Keywords:

Semantic segmentation
Deep learning
Feature extraction
Off-road traversability
Terrain classification

ABSTRACT

This paper introduces a novel approach to dataset annotation across 13 diverse classes for terrain classification. The method is applied to an established set of terrain-related datasets, and a new dataset by the authors' team referred to as CranfieldTerra. These datasets were trained using 18 distinct neural network (NN) architectures, and based on overall test accuracy, precision, recall, and Intersection over Union (IoU) scores, the accuracy of each label, and training time, the effectiveness of these models was evaluated. Furthermore, the methodology has been systematically validated and tested in real-time using a platform called the Husky-A200. This comprehensive evaluation ensures the reliability and accuracy of the methodology under practical conditions.

An innovative real-time switch model is introduced that dynamically selects the most appropriate neural network from the set of pre-trained models based on the calculated environmental density rate and the presence of specific features like 'Person, House, Vegetation Area, and Mud Area' class counts. This approach significantly enhances the adaptability and performance of real-time semantic segmentation in varied environmental conditions, leading to more efficient terrain classification.

1. Introduction

Real-time terrain classification based on semantic segmentation algorithms is a rapidly developing field that has gained significant attention in recent years due to its potential applications in various fields such as robotics, autonomous driving, and surveillance. Semantic segmentation is a computer vision technique that involves dividing an image into multiple segments, each representing a distinct object or region in the image. With the advancements in machine learning algorithms and hardware capabilities, real-time terrain classification based on semantic segmentation has become increasingly feasible, enabling systems to accurately identify and classify various types of terrain in real-time.

A number of studies have been conducted on the topic of semantic segmentation, with significant ones outlined below.

Valada et al. (2020) introduced a multi-modal approach to semantic segmentation using the AdapNet++ architecture and a dataset comprising images from the dataset of Cityscapes (Cordts et al., 2016), Synthia (Ros et al., 2016), SUN RGB-D (Song et al., 2015), ScanNet (Dai et al., 2017), and Freiburg Forest (Valada et al., 2017). They also employed various data augmentation techniques such as cropping, rotation, and scaling to increase the size of the dataset. The Forest

dataset, which included various off-road terrain, was used to validate the study, and the results showed a higher classification accuracy and faster processing time compared to other methods at the time, as depicted in Fig. 1.

A large dataset was created using weakly supervised semantic segmentation for terrain classification and a ground reaction score. To increase the dataset size, the same data augmentation techniques were applied to both RGB images and annotated images with a footpath, which served as input for neural network algorithms. The technique was tested using an ANYmal quadruped robot in an unstructured environment with varying terrain types, weather, and lighting conditions (Wellhausen et al., 2019).

Other researchers have also explored weakly supervised learning approaches in semantic segmentation algorithms. Onozuka et al. proposed a traversability method based on this technique in Onozuka et al. (2021). In their approach, an automatic labelling system was powered by human expertise, with the labelling process involving offline training using trained images and online semantic segmentation. Data augmentation, loss weighting, and preprocessing techniques such as cropping were incorporated into the learning phase. After constructing global and local path planning and utilising semantic segmentation to

* Correspondence to: Old Barns, Bourne End Road, Cranfield, Bedford, MK43 0BD, UK.

E-mail addresses: semihbeycimen@gmail.com (S. Beycimen), D.Ignatyev@cranfield.ac.uk (D. Ignatyev), A.Zolotas@cranfield.ac.uk (A. Zolotas).

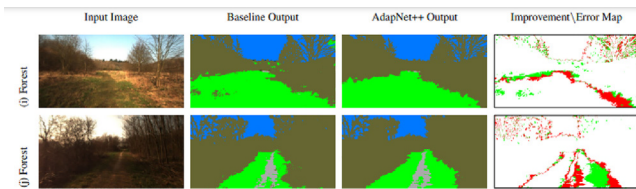


Fig. 1. Semantic segmentation results for the forest dataset (Valada et al., 2020).

identify traversable areas, motion planning was predicted. The study was validated in the real world using a car on structured terrain.

A new method for predicting drivable paths based on weakly-supervised learning has been developed by Barnes et al. (2017). The method uses a large amount of data that has been segmented with a self-supervised labelling technique, allowing a deep learning-based semantic segmentation algorithm to find a reliable path. This approach was tested on the KITTI and Oxford RobotCar datasets under various weather, traffic, and lighting conditions. The authors in Yu et al. (2017) collected the Baidu Driving Dataset (BDD) using three monocular cameras, LIDAR, IMU, and GPS sensors. They used this dataset to predict navigation parameters for lateral and longitudinal control using a deep learning model that includes LSTM. The learning algorithm only used the camera image as input to predict vehicle curvature and speed.

Wolf et al. (2020) developed a hybrid approach that segmented the terrain into sections, such as sky, road, vegetation, grass, and obstacles, allowing for the classification of the terrain and estimation of path traversability. The occupancy map was used to assess traversability, and the study was implemented in both a simulation environment (Unreal Engine 4) and in a real environment with a real-size heavy vehicle (Unimog U5023). Validation trials demonstrated the method's applicability. However, the study's limitation is that the segmented classes are insufficient for off-road terrains.

Based on Soil Property and Object Classification (SPOC), the authors of Rothrock et al. (2016) proposed a method for MARS rover mission, orbital, and ground-based terrain classification that predicts terrain types and features. The study has included two important steps/missions, one of which is the landing site traversability analysis method used to classify the terrain. The method has been applied to 17 terrain types using supervised learning-based semantic segmentation. The labelling of data has been completed manually using web-based labelling tools. In addition, some pre-processing algorithms, such as normalising, darkening, and converting images to bird's-eye images, as well as cropping and masking with RANSAC, were used to create a realistic Mars environment. The study has been confirmed in Columbia Hills, which researchers deemed suitable for the rover landing study. Overall, the method was 90.2% accurate in predicting terrain types such as smooth regolith, dense ridges, rock fields, scarps, and deep sand. However, there was some misclassification, particularly for small rocks, due to data distribution issues.

Another study focuses on improving obstacle detection and semantic mapping in demanding terrain environments. By integrating point-cloud filtering, gradient modelling, and deep learning techniques like Bi-LSTM, the proposed framework effectively identifies hazardous obstacles while ensuring accurate terrain representation. The approach enables safer and more efficient navigation in complex, unstructured scenarios (Song et al., 2023).

A method for producing semantic maps based on the SLAM algorithm has been proposed by Deng et al. (2020). The semantic segmentation algorithm was trained using the RoboCup Rescue-Robot-RGB League's dataset and Inception-v3 learning architecture. Using the depth and RGB images, a SLAM algorithm was then implemented to generate a geometric map. In addition, labelled images have been converted to refined semantic images using a flood-filling algorithm

that employs semantic images and depth data from a stereo camera. To reflect the environment, these semantic images and geometric, semantic, and filtered semantic maps have been generated. This strategy has given reliable analysis by utilising abundant environmental data and more accurate mapping.

In another recent study (Leung et al., 2022), RGB images, point clouds from Lidar and robot motion data have been used to generate the traversability cost map. One of the steps in the method involved converting RGB images into a semantic mask using a supervised semantic segmentation algorithm. The Gated-SCNN (Takikawa et al., 2019), ERFNet (Romera et al., 2017) architectures and RELIS-3D (Jiang et al., 2021), ImageNet (Deng et al., 2009) datasets that have 18 different terrain types have been used to train semantic segmentation algorithms. The main framework and visualisation of the traversability map can be seen in Fig. 2.

Several CNN architectures have been proposed for terrain classification, including Fully Convolutional Networks (FCNs) (Long et al., 2015a), SegNet (Badrinarayanan et al., 2017b), and U-Net (Ronneberger et al., 2015). FCNs are a popular choice due to their ability to handle variable input sizes and produce dense predictions. SegNet is another popular architecture that uses an encoder-decoder architecture with skip connections to improve the accuracy of the segmentation. U-Net is another popular architecture that uses a contracting and expanding path to capture both global and local features of the terrain.

Shojaiee and Baleghi (2023) proposed a multispectral Encoder Fused Atrous Spatial Pyramid Pooling (EFASPP) U-Net deep network to merge the features of the visible and thermal images recorded at night traffic scenes. The proposed network is designed based on the structure of the U-Net, due to its high accuracy and speed of processing by minimising the number of steps in the encoder, as well as no need for large training datasets. Also, a new multispectral dataset is developed in this work for night-time traffic scenes due to the lack of sufficient public datasets in this field.

Arafin et al. (2025) presented a comprehensive survey on terrain classification methods for off-road perception, an area that is highly relevant for autonomous systems operating in challenging environments. Their review carefully categorised different approaches, with particular emphasis on deep learning methods such as Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) for semantic segmentation. They discussed how these techniques enabled pixel-wise classification of terrain types and examined key challenges, including class imbalance, changes in illumination, and the need for robust generalisation in real-world conditions. In addition, the survey provided an extensive overview of publicly available off-road datasets such as RELIS-3D, RUGD, and CaT, which have become fundamental resources for training and evaluating modern terrain classification models.

Shu et al. (2024) provided a detailed overview of Terrain Traversability Evaluation (TE) for autonomous robots, recognising it as a crucial component for scene understanding and navigation. The authors defined TE clearly and distinguished it from related concepts such as ground segmentation and semantic segmentation, with TE considered as a coarse-grained categorisation compared to the fine-grained pixel-level classification of semantic segmentation. Their work organised TE methodologies across several dimensions, including sensor types (cameras, LiDAR, and multisensor fusion), robot types, application scenarios (such as unstructured outdoor terrains), and learning paradigms. The study also summarised common features and computational strategies, highlighting datasets like RELIS-3D and ORFD, and underlined the importance of computational efficiency for real-time applications.

Wang et al. (2025) introduced a method for terrain traversability analysis specifically designed for tracked vehicles in unstructured environments. Their approach generated local traversability maps by combining onboard LiDAR and camera data. A notable contribution of their work was the integration of geometric properties with semantic

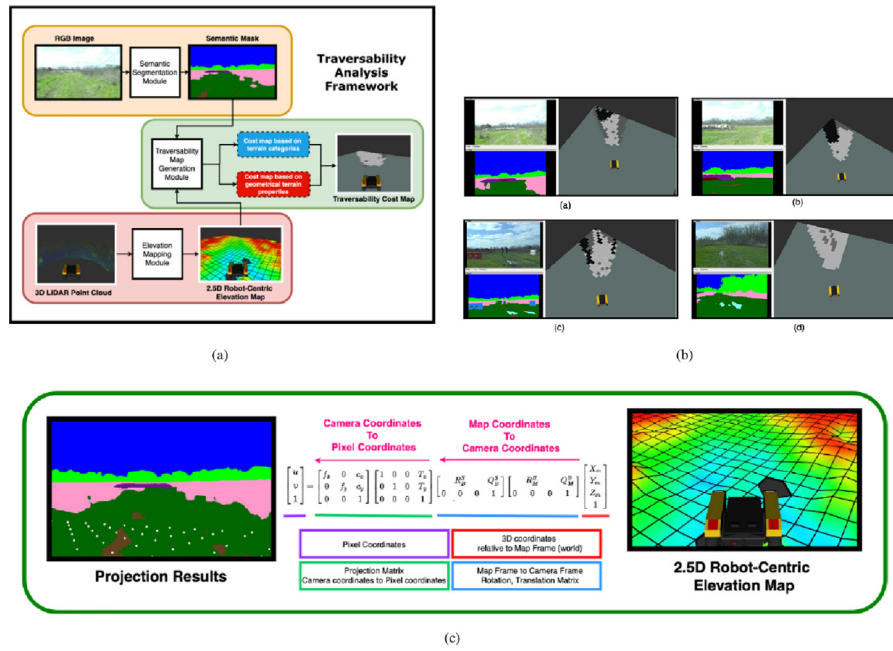


Fig. 2. Figure info: (a) the main framework, (b) Visualisation of traversability cost map (in each image, the top-left is the main image, bottom-left is the semantic mask, and on the right is the terrain traversability cost map. Darker colour shade represents the less traversable area, (c) the projection process (Leung et al., 2022).

terrain information, allowing for more reliable traversability predictions in complex scenarios. The method employed an Open-Vocabulary Segmentation (OVS) module based on the CLIP model (ViT-B/16) to create semantic masks from RGB images, which enhanced generalisability across diverse terrains. Moreover, the authors incorporated the physical capabilities of the vehicle (chassis layer) into the cost calculations, ensuring that the traversability maps were consistent with the robot's actual operational constraints.

Gherold et al. proposed a self-supervised learning method for traversability estimation based on the Cost of Transport (COT) metric, using only RGB-D vision inputs. A key contribution was an automated labelling pipeline leveraging the Segment Anything Model (SAM), which reduced manual effort in dataset creation. They also implemented a real-time inference model on edge devices such as the Nvidia Jetson Orin Nano, demonstrating both efficiency and accuracy in classifying terrain types in real-world scenarios.

Lim et al. (2024) presented a survey clarifying the difference between ground segmentation and traversability estimation. Ground segmentation was described as a preprocessing step for separating ground from non-ground, while traversability estimation was framed as a higher-level process considering robot manoeuvrability, obstacles, and deformable elements. The survey also emphasised the limitations of vision-based sensors, particularly their sensitivity to illumination, which remains a challenge for real-time terrain classification.

Zhang et al. (2025) introduced a sensing and planning system for mobile robots in constrained industrial settings. Their method combined enhanced obstacle parameterisation with a two-stage motion planner (NMPC-DCBF-Convex and DWA-DCBF-Convex), enabling safe navigation in tight spaces. Real-world experiments confirmed the system's ability to avoid obstacles and reach targets with high precision, demonstrating its effectiveness in practical inspection tasks.

Wu et al. (2024) developed an off-road terrain classification algorithm for vehicles using a hybrid 1D CNN-LSTM network. Their method primarily relied on vehicle CAN bus signals and estimated rolling resistance values, rather than vision-based data. The approach classified five common off-road terrain types—namely, asphalt, deep snow, mud, sand, and gravel. Among the three models they tested (LSTM, 1D CNN,

and the combined CNN-LSTM), the hybrid CNN-LSTM achieved the best performance, with an accuracy of 99.2.

Also, several neural network models for semantic segmentation algorithms have been devised. Some of the most recent ones are followed. PlaceNet (Osman et al., 2023) is a plug-and-play-augmented multi-scale deep autoencoder network for scene understanding. M-FasterSeg (Wu et al., 2022) integrates a self-attention module with the neural architecture search (NAS) technique. MGNet (Chang et al., 2022) is a mutual guidance network for few-shot semantic segmentation that improves the discriminative power of class-specific prototypes. Additionally, detailed terrain traversability analysis methods for robotic platforms are reviewed in Sancho-Pradel and Gao (2010), Papadakis (2013) and Beycimen et al. (2023).

1.1. Research and contributions

Existing surveys and related literature have primarily focused on reviewing terrain classification and traversability approaches, highlighting advances, challenges, and available datasets. While these works provide valuable overviews and taxonomies, they are mostly limited to summarising methods and trends. In contrast, our study goes beyond a literature synthesis by introducing a comprehensive multimodal dataset, applying consistent re-labelling across several datasets, benchmarking a wide range of neural network models, and validating the results in real-time conditions. In this sense, our work complements prior surveys by not only discussing the state of the art but also contributing original data, systematic evaluations, and practical validation that strengthen the applicability of semantic segmentation for terrain classification.

This paper aims to explore the current state-of-the-art techniques in real-time terrain classification based on semantic segmentation algorithms and their potential applications in various fields. Furthermore, this paper discusses the challenges associated with implementing such systems and proposes potential solutions to overcome these challenges. The novelty and primary contributions of our study are as follows:

- A comprehensive dataset was collected from a wide range of terrain types and varying lighting conditions, ensuring its robustness for diverse environmental scenarios. The dataset includes

labelled images, 3-D point clouds captured using Lidar and stereo cameras, as well as additional sensor data such as IMU (Inertial Measurement Unit), GPS, odometer readings, and robot description data. This comprehensive dataset not only supports simulation and visualisation efforts (as such large labelled datasets can typically be gathered from simulation environments) but also enables advanced analysis and model development for applications in robotics and autonomous systems. The inclusion of multiple sensor modalities provides a holistic view of the environment, enhancing the dataset's ability to support a variety of research and development needs.

- Several existing datasets, along with our proprietary Cranfield-Terra dataset, were re-labelled into 13 distinct classes using a pixel-based labelling method. This strict re-labelling process ensured high accuracy and consistency over the data. The datasets were trained using 18 unique neural network models, each designed to capture different features and characteristics of the data. This approach allowed for a thorough evaluation of the performance of both the datasets and the neural network models. By comparing the results, we gained valuable insights into the strengths and limitations of each model, as well as the impact of the dataset variations on the model's overall performance, thus facilitating a deeper understanding of their applicability in different scenarios.
- The study is validated in a real-time environment using various pre-trained models, dynamically adjusted based on the environment's density (Real-Time Switch Model). This approach enhances driving performance by adapting to environmental conditions. In more complex environments, a high-accuracy model is employed, while a faster model is used in sparse conditions.

The remaining sections of this paper are as follows. Section 2 discusses on the newly introduced CranfieldTerra dataset. Section 3 presents the methodology followed in this paper in particular, looking at the semantic segmentation process proposed. Results and critical analysis is presented in Section 4. Section 5 concludes the findings of this paper.

2. Datasets for terrain traversability and the CranfieldTerra dataset

In our study, we utilised multiple datasets previously published by other researchers. The Yamaha-CMU Off-Road Dataset (Maturana et al., 2018) contains various classes and was collected from diverse scenes. However, it lacked sufficient data on features such as obstacles, people, houses, and mud classes. Rellis-3D (Jiang et al., 2021) was gathered from vegetational off-road terrain and includes data on mud, people, traversable and untraversable vegetation areas. SEMFIRE forest (Bittner et al., 2021) was collected from three scenarios, and the image quality is not optimal. We utilised this dataset to analyse the effect of poor visibility conditions. Off-Road Terrain Dataset by MAGNUM Research Group (Gresenz et al., 2021) was collected from a dense vegetation area and includes images captured in low-light and low-quality conditions. CaT: CAVS Traversability Dataset (Sharma et al., 2022) was based on a very dense vegetational environment collected in a forest. The Freiburg forest dataset (Valada et al., 2017) includes data from various environmental conditions under different lighting conditions.

The presented datasets are established and largely utilised for terrain classification and traversability studies. Below a few shortcomings are listed:

- Datasets were collected from similar environment types.
- Numerous classes, including persons, buildings, pavement, vehicles, objects, mud areas, extremely rough road areas, and asphalt roads, were insufficiently represented.

- The datasets typically contained data from sparse environments or dense environments dominated by vegetation; however, the number of datasets containing multiple classes was insufficient.
- Some datasets were only published in image format, with no video or other formats such as rosbag format (a compressed file format in the Robot Operating System (ROS)) for validation and testing.
- In addition, the majority of these datasets lack geometric-based data from the environment, such as point clouds and depth information. Therefore, both appearance-based and geometric-based terrain features cannot be extracted using the same dataset.

Moreover, an important aspect of the training of neural network algorithms is that of labelling the datasets and in most cases, the labelling may not be sufficiently detailed to enable more efficient training. This is addressed in the work presented in this paper. Furthermore, our research introduces a novel, comprehensive dataset (referred to as CranfieldTerra) specifically tailored for the task of off-road semantic segmentation, encompassing several diverse environments, lighting conditions, and the presence of shadows. The dataset was diligently curated across several geographical regions with varying environmental densities, providing a robust foundation for training and testing models targeted at off-road traversability analysis. One of the unique aspects of this dataset lies in the detailed categorisation of road types. It includes a wide range of surfaces commonly encountered in off-road and in transition from on to off-road. Specifically, it contains examples of rough roads, asphalt roads, smooth roads, pavement roads, grass roads, stone roads, soil roads, and more as well as mud areas, persons, vehicles and buildings. Each road type offers a distinct set of challenges and characteristics, making the dataset a comprehensive tool for understanding and improving off-road navigation capabilities.

To enable further diversity and capture diverse features in our dataset, we have included images captured under a variety of lighting conditions. The dataset contains images captured at various times of the day, in a variety of weather conditions, and during different times of the year. This allows models trained on our dataset to generalise better to real-world lighting conditions and can help their capability to address diverse lighting situations. Furthermore, the environments encapsulated in our dataset vary in terms of their density, ranging from sparse and wide-open terrains to dense and cluttered landscapes. This diverse environmental density contributes to the creation of a comprehensive dataset, allowing the performance of the model to be evaluated in a variety of scenarios.

The introduced dataset (CranfieldTerra) not only enriches the existing resources for off-road semantic segmentation but also provides a unique perspective due to its carefully curated diversity in road types, lighting conditions, and environmental densities. We believe this dataset will facilitate the development of more effective and robust off-road navigation systems, and serve as a valuable resource for the broader research community. Figs. 4 illustrate the Husky A200 robot used in our study for real-world data collection, serving as the primary platform for acquiring datasets that support experimental evaluation, specifically the CranfieldTerra dataset.

It is worth noting that a few details on the vehicle platform were utilised for the CranfieldTerra data collection. This was performed by the Husky A200, equipped with a mini ITX computer, and an Nvidia GTX 1065 GPU was used for data collection, validation, and testing. Fig. 3 shows the Husky A200 snapshot, including its sensors payload: Velodyne VLP 16 Lidar, 3 Stereolab Zed-2 Stereo Cameras, Duro RTK GPS.

The introduced dataset (CranfieldTerra) together with the existing datasets (Yamaha-CMU Off-Road, Rellis-3D, SEMFIRE Forest, Off-Road Terrain Dataset by MAGNUM Research Group, CaT: CAVS Traversability, Freiburg Forest) were combined to compile a robust dataset containing several classes, scenarios, terrains and object types under varying lighting and camera quality conditions.



Fig. 3. Husky A200 vehicle platform (centre for autonomous & cyber-physical systems, Ground autonomy Lab, Cranfield University).

The extended dataset encompasses both unstructured environments and urban scenarios, including houses, vehicles, people, and asphalt roads. It features a diverse range of terrains, such as smooth roads, rough roads, muddy areas, and asphalt surfaces, among others. The dataset encompasses a wide range of sensor modalities, including GPS, IMU, LiDAR, RGB, and RGB-D data, providing comprehensive environmental perception. Additionally, it includes robot-specific information, such as detailed geometric properties of the robots, joystick input data, and other control-related parameters, enabling advanced simulation and analysis. The dataset for this project is available on GitHub.¹

3. Methodology

In this section, an overview of semantic segmentation has been presented.

3.1. Annotation

In our prior study (Beycimen et al., 2022), we utilised the YAMAHA dataset with their indicated annotation method which contains a number of annotated images of various road types and scenarios. This dataset contains a number of annotated images of various road types and scenarios. However, due to the limited labelling granularity, efficient classification of several images was impacted negatively. To try and mitigate this, the authors in Badrinarayanan et al. (2017a), inspected and aimed to correct manually. Still, the transition area between classes, such as between low-traversable grass and asphalt roads, lacked precision and detail. Also, the dataset contained only eight classes (sky, rough trail, smooth trail, traversable grass, high vegetation, non-traversable low vegetation and obstacle). There are no other classes for primary road terrains besides rough trails and smooth trails. Similar issues that can impact the training of ML approaches

for traversability and terrain mapping where seen in other datasets. Hence, we have followed a more critical re-labelling exercise for all existing datasets (see Section 3.1), including that of CranfieldTerra. Here, a pixel-by-pixel detailed approach was followed by the Matlab Image Labeller Toolbox.²

In this study, the term real-time refers to the ability of the system to process sensory inputs and produce terrain classification outputs within time constraints that allow the robot to react without compromising safe and stable operation. Specifically, the end-to-end response time, measured from data acquisition to model output, was maintained within a latency range that is shorter than the robot's control cycle and decision-making requirements. This ensures that perception results are continuously available to the navigation module at a frequency sufficient for dynamic adjustment of the robot's behaviour in unstructured environments. The measured response times for different models are reported in Table 7, demonstrating that the proposed approach satisfies real-time requirements in practice.

To ensure high-quality and consistent labels across diverse terrains, we adopted a grid-based annotation strategy that combines both fine (small) and coarse (large) grids. The fine grids were applied in areas with complex features — particularly in regions where transmission and structural details were critical — allowing precise pixel-level annotation of small objects and terrain boundaries. In contrast, the larger grids were employed in more homogeneous regions, which reduced redundancy and improved labelling efficiency without compromising accuracy. This dual-grid method provided a balance between annotation precision and scalability, enabling the creation of detailed semantic labels while maintaining consistency across the entire dataset. The proposed strategy differs from conventional annotation practices by explicitly adapting the grid size to local scene complexity, which we believe is a key novelty of our dataset preparation. A wider grid count for the labelling of the clearer deterministic portion of the classes was used, and a more dense grid count for the transition parts between classes to enable more accurate and detailed annotation (see snapshots in Fig. 5(a) and Fig. 5(b)). This was a time-consuming exercise, although work in looking at the more automated way of enabling this is underway.

In total, we re-labelled 1075 images from the YAMAHA dataset, 310 from the Rellis 3-D dataset, 128 from the CaT: CAVS Traversability Dataset, 160 from the Off-Road Terrain dataset by the MAGNUM Research Group, 539 from the Freiburg forest dataset, and 1200 from the newly introduced CranfieldTerra dataset. To summarise, the overall exercise considered 3412 re-labelled data (2212 images from existing datasets in the literature and 1200 from the newly introduced one). To the best knowledge of the authors, the resulting dataset forms the largest dataset in the literature in terms of labelled images with detailed annotation and multiple classes under various lighting, camera quality, and environmental conditions. Figs. 7 and 8 represent some samples of labelled images from existing datasets and our dataset (CranfieldTerra), respectively (see Fig. 6).

3.2. Neural networks

Several models for feature extraction and learning have been evaluated and compared in order to establish more robust models. For feature extraction, the following models were employed: *ResNet50*, *ResNet101*, *ResNet152*, *MobileNetV2*, *InceptionV4*. In addition, the algorithm utilises pre-trained versions of these networks trained on more than a million images from the ImageNet database. The following networks have been used for the primary classification algorithm: *FC – DenseNet56*, *FC – DenseNet67*, *FC – DenseNet103*, *Encoder – Decoder*, *Encoder – Decoder – Skip*, *RefineNet10*, *FRRN – A*, *FRRN – B*, *MobileUNet*, *MobileUNet – Skip*, *PSPNet*, *GCN*, *DeepLabV3*, *DeepLabV3_{plus}*, *AdapNet*, *DenseASPP*, *BiSeNet*, *ERFNet*, and

¹ <https://github.com/semihbeycimen/On-Real-Time-Semantic-Segmentation-with-Comprehensive-Off-Road-Dataset>

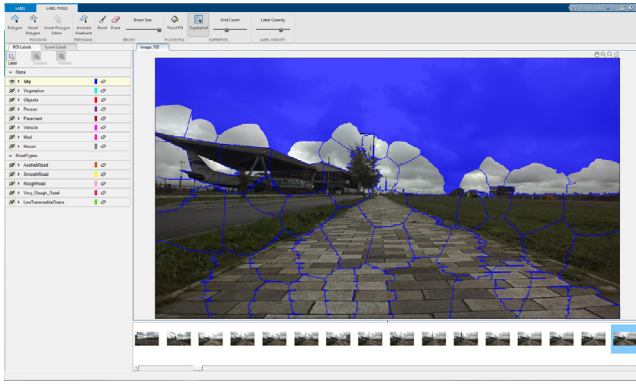
² <https://www.mathworks.com/help/vision/ref/imagelabeler-app.html>



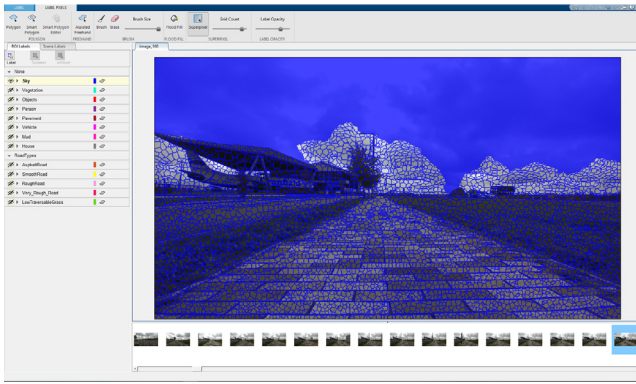
Fig. 4. Examples of image snapshots from CranfieldTerra dataset.

Modified – BiSeNet, respectively (Huang et al., 2017; Long et al., 2015b; Lin et al., 2017b,a; Zhang and Qiao, 2017; Zhao et al., 2017; Kipf and Welling, 2017; Chen et al., 2018; Badrinarayanan et al., 2017a; Chen and Xu, 2018; Yu and Xie, 2018; Romero and Rister, 2017). Some of these NN models used in real-time implementation are explained below.

Full-Resolution Residual Networks (FRRN) model (Pohlen et al., 2017) was proposed to combat resolution loss in conventional CNN-based segmentation models. Due to pooling layers and strided convolutions, the resolution of the feature maps in many deep-learning models decreases as you move deeper into the network. While this helps to extract higher-level semantic features, the loss of fine-grained spatial



(a) Wider Patches Pixel Augmentation for the General Labelling



(b) Denser Patches Augmentation for Transitioning Classes

Fig. 5. Augmentation of the labelling progress.

Class	Red	Green	Blue	Color
Sky	0	0	255	Blue
Vegetation	0	245	196	Cyan
Objects	255	0	0	Red
Person	126	47	142	Purple
Pavement	162	20	47	Brown
Vehicle	255	0	255	Magenta
Mud	255	26	185	Pink
House	128	128	128	Grey
AsphaltRoad	217	83	25	Orange
SmoothRoad	255	255	17	Yellow
Rough Road	246	158	220	Pink
Very Rough Road	246	18	97	Red
LowTraversableGrass	100	212	19	Green

Fig. 6. Colour coding for semantic segmentation classes.

information makes it more difficult to accurately designate class labels at the pixel level. This model possesses the same superior training characteristics as ResNets, but it has two processing streams to address the previously mentioned deficiencies.

- **Residual Stream:** This is a series of convolutional and pooling layers that, similar to conventional CNN architectures, analyse and extract high-level semantic features at multiple scales. Reduced resolution is applied to the feature maps in this broadcast.
- **Pooling Stream:** This stream maintains the original image's resolution throughout the entire network. It integrates the high-level semantic information extracted from the residual stream with the

Table 1

Neural network architecture for ERFNet model.

Layer	Type	out-F	out-Res
1	Downsampler block	16	512×256
2	Downsampler block	64	256×128
3–7	5 x Non-bt-1D	64	256×128
8	Downsampler block	128	128×64
9	Non-bt-1D (dilated 2)	128	128×64
10	Non-bt-1D (dilated 4)	128	128×64
11	Non-bt-1D (dilated 8)	128	128×64
12	Non-bt-1D (dilated 16)	128	128×64
13	Non-bt-1D (dilated 2)	128	128×64
14	Non-bt-1D (dilated 4)	128	128×64
15	Non-bt-1D (dilated 8)	128	128×64
16	Non-bt-1D (dilated 16)	128	128×64
17	Deconvolution (upsampling)	64	256×128
18–19	2 x Non-bt-1D	64	256×128
20	Deconvolution (upsampling)	16	512×256
21–22	2 x Non-bt-1D	16	512×256
23	Deconvolution (upsampling)	C	1024×512

high-resolution feature maps. This enables FRRN to preserve fine-grained spatial information while also considering the semantic context at a high level.

These two streams are linked via “full-resolution residual units” (FRRUs). Each FRRU receives as input the full-resolution feature map from the preceding FRRU (or the initial input image for the first FRRU) and the output of the last unit in the residual stream. It then executes a series of operations, such as convolutions, batch normalisation, ReLU non-linearities, and upsampling, to generate a new high-resolution feature map. FRRN maintains its full resolution throughout the entire network, allowing it to use both high-level semantic information and low-level spatial information for segmentation. Consequently, it can accomplish more precise semantic segmentation than models that operate solely on downsampled feature maps. The neural network architecture can be seen in Fig. 9

Efficient Residual Factorised ConvNet (ERFNet) (Romera et al., 2017) was developed with the intention of achieving a balance between accuracy and computational efficiency, making it appropriate for real-time applications such as self-driving cars. ERFNet's architecture is predicated on factorised convolutions and residual connections. The network consists of encoder and decoder modules, a typical structure for semantic segmentation tasks (see Table 1).

- **Factorised Convolutions:** The central component of ERFNet is factorised convolutions. They reduce the computational complexity of a model by dividing a standard convolution into two simplified operations. A 2D convolution, for instance, can be factorised into two 1D convolutions: one applied vertically and the other horizontally. This method drastically reduces the number of parameters and, consequently, the computation time, making the model more effective and faster.
- **Residual Connections:** ERFNet implements residual connections (or skip connections), a common feature in deep learning models to address the vanishing gradient problem. These connections allow the gradient to travel directly through multiple layers by skipping one or more layers in between, enabling the network to learn identity functions that enhance the model's overall performance.
- **Encoder-Decoder Structure:** The encoder component of ERFNet reduces the spatial dimension of the input image while increasing the depth of feature maps, thereby capturing the scene's semantic information. The decoder then progressively recovers the spatial information, ensuring that the output image has the same dimensions as the original, and assigns a class label to each pixel.
- **Non-bottleneck-1D:** The use of so-called “non-bottleneck-1D” modules is another notable aspect of the ERFNet architecture.

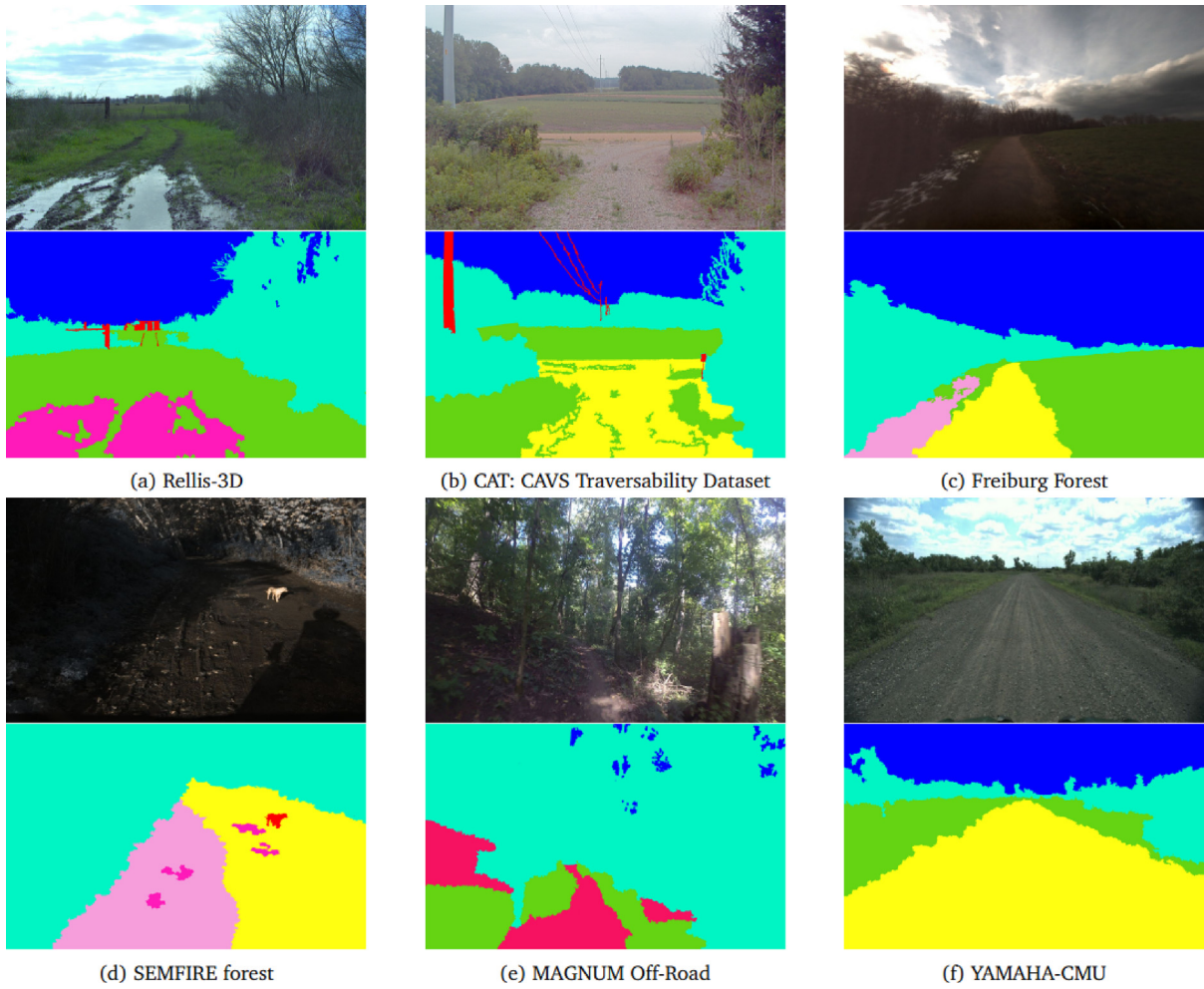


Fig. 7. Example image snapshots and data labelling from existing datasets in the literature of terrain mapping.

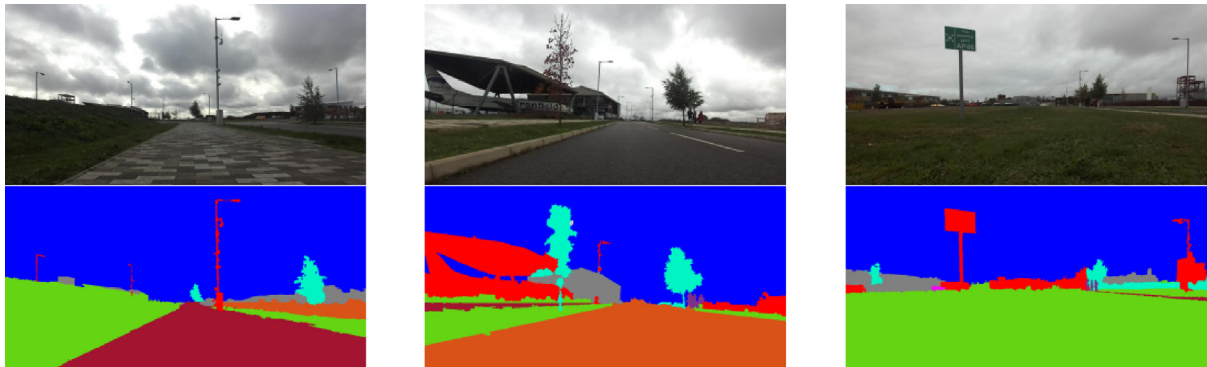


Fig. 8. Data labelling of CranfieldTerra dataset.

These modules consist of a series of 1D convolutions and are designed to maintain the input's original resolution throughout the network, allowing it to capture more detailed spatial information and enhancing the segmentation's quality.

The densely connected Atrous Spatial Pyramid Pooling (DenseASPP) (Yang et al., 2018) model was created to address the difficulty of image scale in semantic segmentation tasks. It builds on Atrous Spatial Pyramid Pooling (ASPP), which employs parallel atrous convolutions with various dilation rates to capture multi-scale context. The central concept underlying DenseASPP is the dense connection of a collection

of atypical convolutional layers. Atrous convolution, also known as ensilated convolution, controls the field of view of the filters, enabling the model to capture spatial information at various scales. Using atrous convolutions with varying dilation rates, the model is able to effectively extract image features at different scales. The “dense” component of DenseASPP denotes that each atrous convolutional layer receives inputs from all preceding layers, not just its immediate predecessor. This makes the model's information flow more efficient, enabling it to generate multiscale features that cover a larger scale range more densely, hence the name “DenseASPP”. This model is especially useful for applications where high-resolution comprehension is essential, as

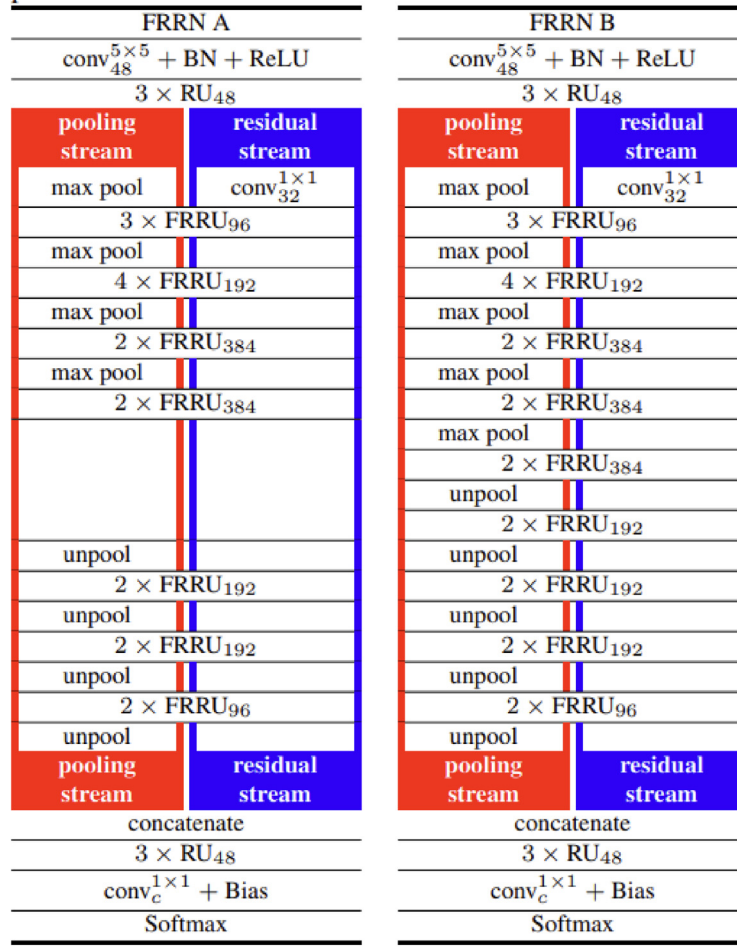


Fig. 9. Neural network architecture for FRRN A and B models.

it can capture details more efficiently without substantially increasing the model's size.

3.3. Real-time switch model

The semantic segmentation framework was built on labelled data re-processed into 13 distinct classes using a pixel-based labelling method. These labelled datasets were initially trained on 18 neural network models, allowing for comprehensive experimentation. However, only four models — FRRN-B, DenseASPP, and ERFNet — were ultimately selected for the Real-Time Switch Method due to their superior performance metrics in terms of accuracy and response time.

In this study, we introduced the *Real-Time Switch Method* to dynamically adapt the performance of our semantic segmentation models based on the density rate of the environment. The proposed methodology ensures an optimal balance between accuracy rate and response time by assigning specific models to varying levels of environmental complexity. The model selection criteria and their designated use cases are described as follows:

3.3.1. FRRN-B model for high-density environments

The FRRN-B model was specifically designated for scenarios characterised by challenging environments, where the density rate exceeded 50%. Such scenarios typically include cluttered spaces with a high number of classifiable elements requiring robust semantic segmentation. The FRRN-B model's architecture, designed with a strong focus on feature refinement and residual connections, makes it particularly adept at processing complex scenes. Its ability to maintain high accuracy in densely populated environments ensures precise segmentation, even when computational demands are high.

3.3.2. DenseASPP and ERFNet models for moderate-density environments

For environments classified as *normal*, where the density rate ranged between 25% and 50%, the DenseASPP and ERFNet models were employed. This range represents moderately cluttered environments requiring a balance between computational efficiency and segmentation precision.

- **DenseASPP:** With its densely connected atrous spatial pyramid pooling, the DenseASPP model is effective in capturing multi-scale context while preserving computational efficiency.
- **ERFNet:** Known for its lightweight architecture, the ERFNet model complements DenseASPP by offering faster processing times while maintaining acceptable accuracy.

The combination of these models ensures reliable segmentation in moderately complex environments, striking an effective compromise between speed and precision.

3.3.3. DenseASPP model for low-density environments

In sparse environments, where the density rate fell below 25%, the DenseASPP model alone was deemed sufficient. Sparse environments typically consist of fewer objects of interest, reducing the need for computationally intensive models. The DenseASPP model's efficient design allows it to deliver quick processing times without a significant trade-off in accuracy, making it an ideal choice for real-time applications in less demanding scenarios.

This dynamic model-switching strategy ensures that computational resources are utilised efficiently while maintaining high segmentation accuracy across diverse environmental conditions. The Real-Time

Switch Method enables the system to dynamically respond to the complexity of the surrounding environment, thereby optimising real-time performance and ensuring robust adaptability.

4. Results

In this section, we describe the training, testing, and underlying training architecture details.

The YAMAHA dataset was initially trained on a supercomputer equipped with a V100 GPU card and two Intel E5-2620 v4 (Broadwell) processors. The input images for the learning model had a pixel height and width of 512 and 1024, respectively. The training parameters were: batch size 1, Root Mean Squared Propagation (RMSPropOptimizer) with a decay of 0.995 and a learning rate of 0.001.

4.1. Model optimisation step

The training parameters described in the paper were determined through a systematic model optimisation process to ensure optimal performance as follows:

4.1.1. Batch size selection

Various batch sizes (1, 2, 4, and 8) were tested to evaluate the trade-offs between computational efficiency and model performance. Due to the high resolution of the input data and hardware limitations, a batch size of 1 was selected as the most suitable, ensuring efficient memory usage without compromising convergence stability.

4.1.2. Optimiser choice

RMSProp was selected after comparing it with alternative optimisers such as Adam and SGD. RMSProp demonstrated superior convergence behaviour on our dataset, especially given the high variability in off-road environments based on the literature. The decay parameter was set to 0.995 based on literature recommendations and experimental validation. Specifically, values ranging from 0.9 to 0.999 were tested, and the selected value provided a stable optimisation trajectory without significant oscillations.

4.1.3. Learning rate tuning

A grid search was conducted to optimise the learning rate. Values in the range of 10^{-5} to 10^{-2} were evaluated incrementally. A learning rate of 0.001 was chosen as it provided a good balance between convergence speed and stability, avoiding issues such as vanishing or exploding gradients. Higher learning rates led to unstable training, while lower rates significantly slowed down convergence.

4.1.4. Validation-based fine-tuning

The parameters were fine-tuned based on the model's performance on the validation set. Metrics such as Intersection over Union (IoU) and pixel-wise accuracy were monitored during this phase to evaluate the impact of each parameter adjustment.

4.1.5. Early stopping and regularisation

To prevent overfitting during training, early stopping was employed, triggered by monitoring the validation loss. This ensured that the selected parameters were effective across the diverse density configurations in the dataset.

This optimisation process ensured robust model performance while minimising unnecessary computational overhead.

4.2. Overall accuracy results and training time for network models

We trained the dataset in 200 epochs under the aforementioned conditions using the 18 NN models. In our previous paper, we analysed five CNN-based feature extraction (frontend) models (ResNet-50, ResNet-101, ResNet-152, MobileNetV2, and InceptionV4). The ResNet-152 model was used as a feature extractor in this study due to its accuracy, training time, and computational demand. 70% of the dataset was used for training, 15% for validation, and 15% for testing.

Throughout the training process, we extracted several results to .csv files, including Validation precision, recall, F1 and IoU scores, as well as average validation accuracies per class for each epoch. Thus, we chose the best-trained model based on the highest validation accuracy and IoU score. Then, we applied this trained model to the test dataset, which had not been used in any training step/phase. The Tables 3 and 4 contain our test results for the YAMAHA dataset along with our detailed labelling. In addition, the results from our previous paper are included in Table 2 in order to evaluate the impact of our labelling model. As shown in Tables 2 and 3, our annotation model improved accuracy by an average of 10%. However, some class accuracy results are close to 1 or unreliable due to the distribution of the dataset and insufficient label quantity, as can be seen in the columns from "Person" to "Asphalt Road" in Table 4.

Therefore, we generated a new dataset using the existing six datasets as well as our new dataset, which was compiled with the limitations of other datasets. After a long and complex labelling process, all NN models were re-trained using this combined dataset under identical conditions. Once more, the dataset is divided into 70% for training, 15% for validation, and 15% for testing. We manually added images from all classes and terrain types to the train, validation and test sub-datasets, which were hand-picked (to set better data distribution). Therefore, we attempted to provide balanced training, validation, and test data from each class. Both 5 and 6 contain the combined dataset's results. Due to a larger and more evenly distributed data set, a more reliable model has been developed, which yielded positive results for the classes of mud, obstacle, person, and vehicle, despite a slight decrease in average accuracy (see Tables 5 and 6).

4.3. Real-time implementation

For our real-time implementation, we rigorously evaluated and subsequently employed four distinct neural network models known for their precision and expedited processing and response capabilities. The chosen models—FRRN-B, DenseASPP, BiSeNet, and ERFNet were deployed for empirical testing within various terrains (Fig. 11), encompassing both off-road and semi-off-road areas within the confines of Cranfield Campus. The evaluation process involved selecting the top three models with the highest accuracy rates and the model with the fastest inference time. These models were chosen based on their performance metrics during validation, ensuring a balance between precision and computational efficiency for varying environmental conditions. The accuracy rate and response time for each model are illustrated in Table 7.

Model selection was meticulously determined by the calculated 'environment density rate.' This critical metric was derived by dividing the total pixel count of each image by the aggregated count of pixels attributed to key environmental features—namely "Vegetation", "Objects", "Person", "Vehicle", "Mud" and "House" classes.

The subsequent selection criteria for the models were as follows and can be seen in Fig. 10:

FRRN-B Model: This model was designated for scenarios characterised by exceedingly dense environments, where the density rate exceeded %50. The FRRN-B's robust architecture makes it adept at handling complex scenes with a high density of classifiable elements.

Table 2

Comparison of semantic segmentation NN methods for YAMAHA dataset with original labelling.

Network model	No. of total parameters	Training time (hrs)	Overall accuracy	Trvsble grass	Rough trail	Smooth trail	Obstacle	Truck (Bckgnd)	Non Trvsble	Sky	High Veg.
FC-DenseNet56	1 365 632	08:22:28	79.5	72	70	82	73	72	68	97	94
FC-DenseNet67	3 442 536	08:17:54	77.6	76	69	74	76	75	68	88	93
FC-DenseNet103	9 269 784	08:30:12	78.0	59	70	84	73	76	71	96	94
Encoder-Decoder (based on SegNet)	34 968 264	07:33:22	78.8	69	67	81	73	79	74	94	93
Encoder-Decoder (with Skip)	34 968 264	07:36:48	79.3	73	74	83	72	78	65	95	95
Refinet	85 684 616	07:58:30	75.5	67	70	80	75	77	68	93	96
FRRN-A	17 740 088	10:50:48	78.2	65	65	80	74	78	69	96	93
FRRN-B	24 748 984	10:52:44	79.5	72	68	83	75	78	66	95	96
Mobile-Unet	8 872 392	07:56:05	79.2	73	69	84	73	76	67	91	95
MobileUNet-Skip	8 872 392	07:56:47	79.4	69	64	92	75	77	68	95	95
PSPNet	56 000 328	05:44:21	74.7	61	71	71	72	78	69	95	92
GCN	42 994 642	05:38:16	75.2	63	66	86	74	76	69	95	94
DeepLaB-v3	46 661 256	04:34:23	74.3	76	71	78	72	72	66	95	91
DeepLaB-v3+	47 955 128	04:43:30	78.7	72	71	75	76	79	68	97	97
AdapNet	21 069 187	05:35:19	76.5	60	65	86	73	73	70	94	92
DenseASPP	43 767 112	04:23:05	77.4	56	63	85	74	79	73	92	93
BiSeNet	47 564 712	05:45:51	78.9	64	64	88	77	78	69	96	95
Modified BiSeNet	44 476 424	06:38:28	79.1	68	70	87	79	79	70	96	96

Note: Values in columns 4-9 given in %; Trvsble = Traversable; Bckgnd = Background; Veg. = Vegetation.

Table 3

Comparison of general parameters for semantic segmentation NN methods with YAMAHA dataset with our detailed labelling.

Network models	Train time (100 Epochs)	Average accuracy	Precision	Recall	F1 score	IoU score
FRRN-A	12:10:34	90.7	90.1	88.7	88.8	72.1
BiSeNet	09:22:43	88.3	90.0	88.3	88.5	71.6
FRRN-B	12:29:52	88.1	90.2	88.0	88.5	71.1
ERFNet	08:12:57	87.9	89.4	88.1	87.9	71.1
Encoder-Decoder with skip	08:57:03	87.6	88.9	87.6	87.7	70.9
DenseASPP	04:50:26	87.9	89.6	87.9	88.2	70.8
MobileUNet-Skip	17:28:53	88.2	90.0	88.2	88.7	70.8
GCN	09:45:56	87.9	89.4	87.9	88.0	70.7
FC-DenseNet56	11:56:33	88.2	90.0	88.2	88.5	70.7
PSPNet	07:29:51	87.1	88.3	87.1	87.2	70.5
Encoder-Decoder	07:34:56	87.1	88.1	87.1	86.9	70.4
DeepLaB-v3+	07:32:50	87.6	89.1	87.6	87.9	70.1
AdapNet	07:33:01	86.7	88.7	86.7	86.8	69.9
FC-DenseNet103	15:39:51	88.4	90.4	88.1	88.6	69.9
Mobile-Unet	17:25:23	87.3	88.8	87.3	87.5	69.8
FC-DenseNet67	15:49:14	87.9	90.4	89.9	88.6	69.6
Refinet	10:49:39	86.8	88.0	86.8	86.7	69.5
DeepLaB-v3	07:32:56	85.6	87.3	85.6	85.8	67.6

DenseASPP and ERFNet Models: These models were utilised under conditions classified as 'normal'—where the environment's density rate ranged from %25 to %50. This intermediate range is indicative of moderately cluttered environments that require a balance between precision and computational efficiency.

DenseASPP Model: In contrast, for sparse environments where the density rate fell below %25, the DenseASPP model alone was deemed sufficient. Its configuration is particularly suited for areas with fewer objects of interest, allowing for quicker processing times without a significant trade-off in accuracy.

The thorough evaluation conducted during our experimentation phase showcased the efficiency of our classification model. It produced precise predictions, demonstrating high performance across a spectrum of environmental conditions. This success can be attributed to the refined labelling techniques and the implementation of a 'real-time switch model'—a strategy that dynamically selects from these four pre-trained models based on the evaluated environmental density.

As shown in Fig. 12(a), the real-time tests were conducted with the Husky A200 robot navigating through a variety of off-road terrains while performing semantic segmentation using the DenseASPP model,

Table 4

Comparison of parameters per class for semantic segmentation NN methods with YAMAHA dataset with our detailed labelling.

Network models	Sky	Vegetation	Objects	Person	Pavement	Vehicle	Mud	House	Asphalt road	Smooth road	Rough road	Very rough road	Low traversable grass
FRRN-A	91.3	93.3	91.5	93.3	100.0	93.7	96.6	94.7	99.1	95.9	89.2	94.0	78.0
BiSeNet	90.8	92.8	91.0	93.3	100.0	93.3	96.6	93.8	96.8	96	84.2	91.6	81.6
FRRN-B	87.4	93.7	93.2	93.3	100.0	93.5	96.6	95.2	98.8	95.8	83.5	91.1	76.6
ERFNet	91.1	90.4	90.5	92.1	100.0	93.3	96.6	93.5	98.1	95.1	86.1	90.0	85.6
Encoder-Decoder with skip	91.8	91.9	93.1	93.5	100.0	93.3	96.6	94.4	99.1	95.4	83.2	92.4	73.8
DenseASPP	88.1	93.5	90.4	93.4	100.0	93.9	96.6	95.1	98.8	95.2	83.3	91.9	78.9
MobileUNet-Skip	85.1	93.2	92.6	93.3	100.0	93.3	96.6	94.2	99.1	95.8	84.1	92.5	75.2
GCN	83.4	94.0	90.0	93.3	100.0	93.6	96.6	93.3	98.2	95.4	83.9	93.2	80.5
FC-	88.2	92.1	90.0	93.6	100.0	93.3	96.6	93.4	98.8	96.0	83.3	91.6	78.3
DenseNet56													
PSPNet	89.2	92.7	90.9	93.4	100.0	93.8	96.6	93.3	98.4	96.1	83.6	94.5	70.6
Encoder-Decoder	88.2	90.2	90.1	93.5	100.0	93.4	96.6	94.7	98.6	96.1	82.2	92.4	79.4
DeepLab-v3+	88.9	93.3	90.3	93.3	100.0	93.4	96.6	94.0	98.4	95.1	82.6	93.3	72.9
AdapNet	89.1	91.1	90.6	93.9	100.0	93.3	96.6	93.3	97.9	94.2	86.7	91.8	75.4
FC-	83.3	93.5	90.2	93.3	100.0	93.3	96.6	93.3	98.1	96	79.9	93.6	75.9
DenseNet103													
Mobile-Unet	84.6	91.7	90.5	94.1	100.0	93.3	96.6	96.1	99.2	95.8	81.4	93.4	75.1
FC-	89.6	91.9	90.1	93.4	100.0	93.1	96.6	93.3	99.1	96.3	79.8	92.3	72.6
DenseNet67													
Refinet	87.6	93.0	90.1	93.3	100.0	93.3	96.6	93.3	97.5	96	82.7	94.1	74.5
DeepLab-v3	86.8	91.9	92.1	93.3	100.0	93.6	96.6	93.7	97.6	96.7	80.6	90.9	71.7

Table 5

Comparison of general parameters for semantic segmentation NN methods with combined dataset with our detailed labelling.

Network models	Train timetabbr (100 Epochs)	Average tabbr accuracy	Precision	Recall	F1 score	IoU score
FRRN-B	21:06:11	90.1	91.6	90.9	90.8	68.8
FRRN-A	20:33:44	89.9	90.3	89.9	89.5	68.7
BiSeNet	15:51:44	89.7	90.1	89.7	89.2	68.5
ERFNet	13:46:19	88.6	89.5	89.1	88.4	67.9
Encoder-Decoder with skip	15:04:25	90.0	90.0	90.0	89.7	66.6
MobileUNet-Skip	29:40:48	87.8	87.6	87.8	87.1	64.2
Encoder-Decoder	14:58:23	89.2	89.6	89.2	88.8	64.1
PSPNet	13:39:55	88.2	89.5	88.2	88.1	63.4
AdapNet	12:45:16	87.7	88.8	87.7	87.5	62.1
GCN	16:32:23	87.1	87.2	87.2	86.3	61.8
DenseASPP	7:54:34	88.4	89.2	88.4	87.8	61.7
FC-	20:21:55	88.3	89.3	88.3	88.1	61.4
DenseNet56						
Mobile-Unet	29:53:01	87.1	88.5	87.3	86.9	60.5
DeepLab-v3+	12:28:15	86.7	87.4	86.7	85.9	60.5
Refinet	18:32:20	86.1	86.9	86.1	85.4	59.1
FC-	26:36:08	86.9	88.1	86.9	86.4	58.7
DenseNet103						
FC-	23:31:42	87.1	88.3	87.1	86.9	58.6
DenseNet67						
DeepLab-v3	12:21:11	81.4	83.8	81.4	80.1	53.0

which was used its reliable performance and faster inference speed. The system was able to successfully process and segment the environment in real time, with the Husky moving autonomously. The robot demonstrated the ability to consistently produce accurate segmentations despite the dynamic and challenging real-world conditions.

The system performed reliably, with smooth operation and minimal latency, as supported by the real-time accuracy and speed metrics reported in Table 7. These quantitative results confirm the robustness and practicality of the proposed approach in real-world conditions, where real-time processing is essential for safe and efficient autonomous

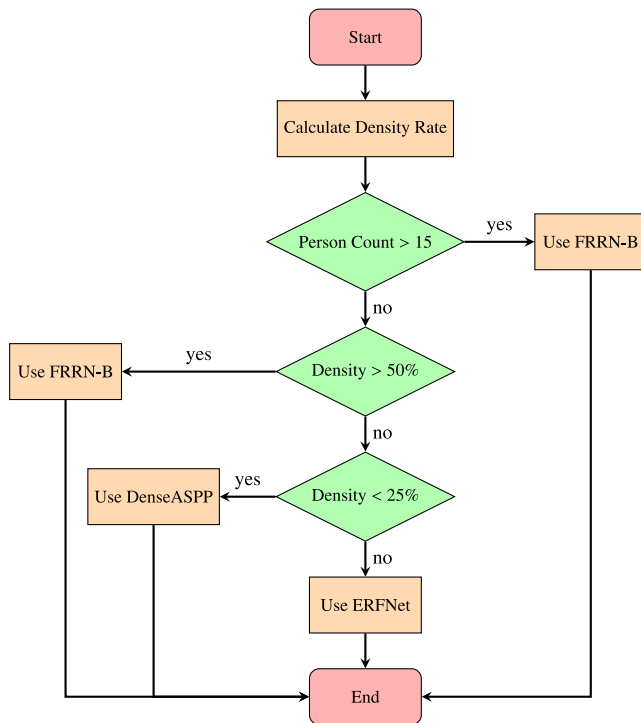
navigation. Additionally, Fig. 12(b) provides a top-view image captured during the validation phase.

Shortcomings, such as detecting mud or hole classes, can be addressed using geometry-based methods or aerial images. The radar-based methods has been examined in a survey by Ersu, Petlenkov, and Janson (Ersu et al., 2024), which provides a comprehensive analysis of advancements in radar technology for UGVs operating in challenging environments. The survey highlights how the integration of radar with machine learning and sensor fusion significantly enhances perception, mapping, and navigation capabilities in diverse terrains, such as foggy, rainy environments, and dense vegetation.

Table 6

Comparison of parameters per class for semantic segmentation NN methods with combined dataset with our detailed labelling.

Network models	Sky	Vegetation	Objects	Person	Pavement	Vehicle	Mud	House	Asphalt road	Smooth road	Rough road	Very rough road	Low traversable grass
FRRN-B	90.8	88.7	81.3	82.3	93.6	83.6	66.9	88.5	95.4	93.9	86.7	93.4	85.4
FRRN-A	91.2	84.2	81.3	83.6	93.9	86.5	67.9	89.6	95.3	92.1	86.5	93.2	86.1
BiSeNet	89.8	85.3	82.8	85.6	93.7	89.2	68.8	87.9	95.5	82.2	85.1	89.1	87.9
ERFNet	90.1	84.2	79.8	83.4	92.4	84.5	67.9	87.1	95.1	84.6	84.6	89.0	88.9
Encoder-Decoder with skip	90.6	84.2	77.6	82.1	93.1	86.6	63.9	73.8	94.4	97.3	86.9	87.8	82.3
MobileUNet-Skip	90.4	81.1	79.7	82.9	93.1	83.1	62.5	85.5	94.3	92.9	82.5	87.2	85.7
Encoder-Decoder	89.9	77.7	73.9	79.3	92.6	79.6	66.8	85.9	94.5	93.6	85.4	90.6	83.9
PSPNet	91.1	83.0	74.6	79.3	88.2	79.8	61.8	81.1	94.9	91.7	83.8	91.1	84.1
AdapNet	92.7	77.6	71.4	79.3	90.5	84.5	56.2	84.4	91.3	82.4	84.8	91.2	81.7
GCN	89.7	81.9	76.3	80.1	90.1	78.5	65.1	78.8	94.2	91.5	81.5	89.0	80.5
DenseASPP	86.5	90.1	62.3	80.8	89.9	79.5	62.0	78.4	94.8	94.4	83.3	90.3	80.8
FC-	92.3	79.9	69.9	80.5	91.4	73.5	69.2	78.8	94.3	91.7	81.9	89.9	80.2
DenseNet56	87.4	80.8	73.7	79.4	89.9	83.6	62.7	86.9	94.9	84.8	86.7	88.6	79.6
Mobile-Unet	92.3	74.9	69.8	79.8	92.0	78.1	59.8	87.6	91.1	93.7	78.8	86.2	88.9
DeepLaB-v3+	91.2	75.9	74.4	79.3	89.7	82.9	58.4	75.6	94.1	89.0	81.8	86.5	83.5
Refinet	87.6	74.5	68.9	74.3	91.2	74.1	59.8	81.1	92.9	92.7	76.9	89.1	89.2
FC-	89.6	77.1	74.1	79.4	90.2	71.7	64.1	72.5	93.2	88.5	83.1	94.7	87.6
DenseNet103	82.7	88.4	64.4	79.5	90.7	72.1	61.4	68.5	90.5	84.6	85.1	85.8	81.9
FC-													
DenseNet67													
DeepLaB-v3													

**Fig. 10.** Flowchart illustrating the model selection process based on environmental density rate.

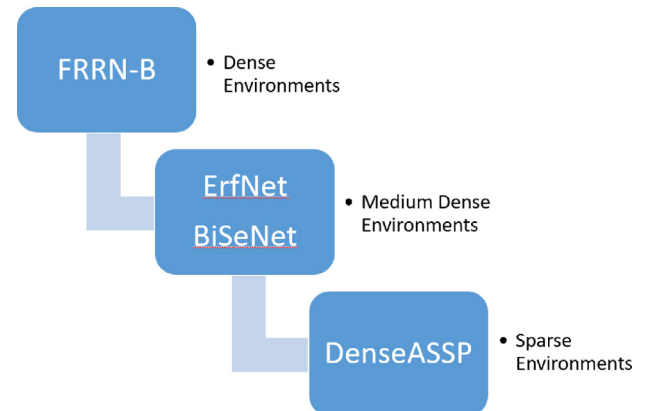
The code, testing progress, dataset, and videos for this project are available on GitHub.³

³ <https://github.com/semihbeycimen/On-Real-Time-Semantic-Segmentation-with-Comprehensive-Off-Road-Dataset>

Table 7

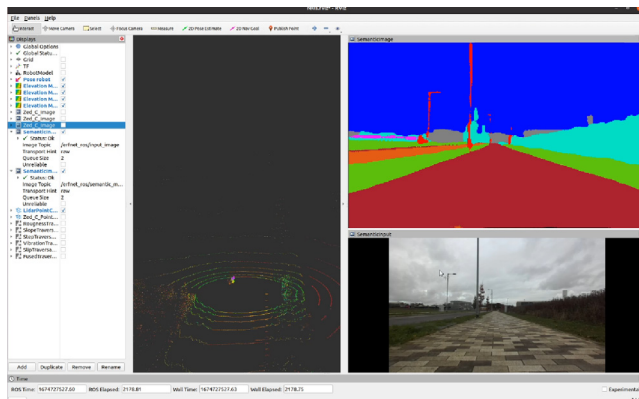
Accuracy Rate and Response Time of Chosen NN Models (Mini ITX Computer with Nvidia GTX1060 GPU).

Network models	Average test Acc.	Response time
FRRN-B	90.7	0.75s
BiSeNet	88.1	0.58s
ERFNet	87.9	0.49s
DenseASPP	87.1	0.25s

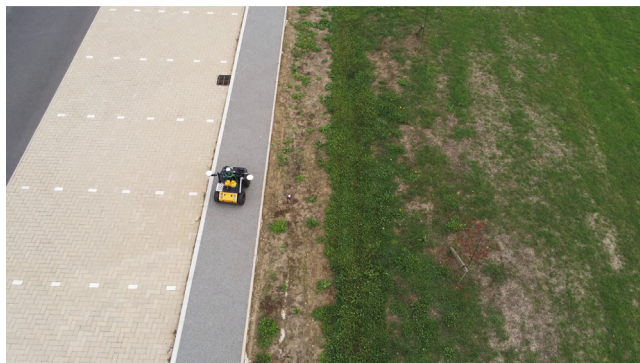
**Fig. 11.** Real-time switch model.

5. Conclusion

In this study, a comprehensive, detailed, pixel-based annotation methodology is outlined and applied to multiple datasets published by other researchers, as well as our extensive dataset. After training the data with the 18 neural networks (NN), the test accuracy, precision, recall, F1 and IoU scores, and average accuracy have been compared for each label. Despite the fact that some models produce more precise outcomes, their response times are slower. Alternatively, some models had a lower rate of accuracy but were faster. For the real-time implementation, we used the FRRN-B model for dense environments



(a) A Semantic Segmentation Result During Real-Time Validation



(b) The Top-View Image During Real-Time Validation and Testing

Fig. 12. Real-time semantic segmentation based terrain classification.

that require greater precision, the DenseASPP model for sparse environments that should respond faster, and the BiSeNet and ErfNet models for the general case. This study demonstrates that by perceiving the environment, multi-model training models enable a more robust real-time implementation of these switch models. Furthermore, the real-time switch model dynamically selects the most suitable neural network based on the environmental density and computational requirements, ensuring that the system can maintain high performance while minimising response times. This approach allows the system to adapt seamlessly to varying conditions, providing both accuracy and efficiency in real-time applications. Consequently, our method was able to classify the terrain accurately in real-time using a real vehicle platform (Husky A200). It was also possible to segment a road, object, vehicle, or house in the distance using this extensive method.

CRediT authorship contribution statement

Semih Beycimen: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Ali Art:** Funding acquisition, Formal analysis, Data curation, Conceptualization. **Mehmet Unal:** Funding acquisition, Formal analysis, Data curation, Conceptualization. **Dmitry Ignatyev:** Writing – review & editing, Supervision, Resources, Conceptualization. **Argyrios Zolotas:** Writing – review & editing, Supervision, Investigation, Conceptualization.

Declaration of competing 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.

Acknowledgment

The first author acknowledges the Republic of Turkey, Ministry of National Education (1416-YLSY), for supporting the studies under the PhD scholarship ref. U9BYTAB2LDGA7LK.

Data availability

Data will be made available on request.

References

- Arafin, T., Hosen, A., Najdovski, Z., Wei, L., Rokonzaman, M., Johnstone, M., 2025. Advances and trends in terrain classification methods for off-road perception. *J. Field Robot.*
- Badrinarayanan, V., Kendall, A., Cipolla, R., 2017a. AdapNet: Adaptive convolutional neural networks for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR*, pp. 24–32, URL <https://arxiv.org/abs/1611.09960>.
- Badrinarayanan, V., Kendall, A., Cipolla, R., 2017b. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 39 (12), 2481–2495.
- Barnes, D., Maddern, W., Posner, I., 2017. Find your own way: Weakly-supervised segmentation of path proposals for urban autonomy. In: *2017 IEEE International Conference on Robotics and Automation. ICRA, IEEE*, pp. 203–210.
- Beycimen, S., Ignatyev, D., Zolotas, A., 2022. A comparison study of neural network-based semantic segmentation for off-road traversability. *UKRAS22 Conf.* 72–73.
- Beycimen, S., Ignatyev, D., Zolotas, A., 2023. A comprehensive survey of unmanned ground vehicle terrain traversability for unstructured environments and sensor technology insights. *Eng. Sci. Technol. Int. J.* 47, 101457.
- Bittner, D., Andrada, M., Portugal, D., Ferreira, J., 2021. SEMFIRE forest dataset for semantic segmentation and data augmentation.
- Chang, Z., Lu, Y., Wang, X., Ran, X., 2022. MGNet: Mutual-guidance network for few-shot semantic segmentation. *Eng. Appl. Artif. Intell.* 116, 105431.
- Chen, X., Xu, H., 2018. Denseaspp for semantic segmentation in street scenes. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR*, URL <https://arxiv.org/abs/1806.01171>.
- Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. DeepLabv3+: Exploring the potential of efficient feature learning for semantic image segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR*, pp. 6952–6961, URL <https://arxiv.org/abs/1802.02611>.
- 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.
- Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M., 2017. ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 5828–5839.
- 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. IEEE*, pp. 248–255.
- Deng, W., Huang, K., Chen, X., Zhou, Z., Shi, C., Guo, R., Zhang, H., 2020. RGB-D Based Semantic SLAM Framework for Rescue Robot. *Proc. - 2020 Chin. Autom. Congr. CAC 2020* 6023–6028. <http://dx.doi.org/10.1109/CAC51589.2020.9326866>.
- Ersü, C., Petlenkov, E., Janson, K., 2024. A systematic review of cutting-edge radar technologies: Applications for unmanned ground vehicles (UGVs). *Sensors* 24 (23), 7807.
- Gherold, V., Mandralis, I., Sihite, E., Gharib, M., Salagame, A., Ramezani, A., Self-supervised cost of transport estimation for multimodal path planning. *IEEE Robot. Autom. Lett.*
- Gresenz, G., White, J., Schmidt, D.C., 2021. An off-road terrain dataset including images labeled with measures of terrain roughness. In: *2021 IEEE International Conference on Autonomous Systems. ICAS, IEEE*, pp. 1–5.
- Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR*, pp. 4700–4708, URL <https://arxiv.org/abs/1608.06993>.
- 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, IEEE*, pp. 1110–1116.
- Kipf, T.N., Welling, M., 2017. Semi-supervised classification with graph convolutional networks. In: *Proceedings of the International Conference on Learning Representations. ICLR*, URL <https://arxiv.org/abs/1609.02907>.
- Leung, T.H.Y., Ignatyev, D., Zolotas, A., 2022. Hybrid Terrain Traversability Analysis in Off-road Environments. In: *2022 8th International Conference on Automation, Robotics and Applications. ICARA 2022*.

- Lim, H., Oh, M., Lee, S., Ahn, S., Myung, H., 2024. Similar but different: A survey of ground segmentation and traversability estimation for terrestrial robots. *Int. J. Control. Autom. Syst.* 22 (2), 347–359. <http://dx.doi.org/10.1007/s12555-023-0826-4>.
- Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., 2017a. Feature pyramid networks for object detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR*, pp. 2117–2125, URL <https://arxiv.org/abs/1612.03100>.
- Lin, G., Milan, A., Shen, C., Reid, I., 2017b. RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR*, pp. 536–545, URL <https://arxiv.org/abs/1611.06612>.
- Long, J., Shelhamer, E., Darrell, T., 2015a. Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp.* 3431–3440.
- Long, J., Shelhamer, E., Darrell, T., 2015b. Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR*, pp. 3431–3440, URL <https://arxiv.org/abs/1411.4038>.
- Maturana, D., Chou, P.W., Uenoyama, M., Scherer, S., 2018. Real-time semantic mapping for autonomous off-road navigation. In: *Field and Service Robotics. Springer*, pp. 335–350.
- Onozuka, Y., Matsumi, R., Shino, M., 2021. Weakly-supervised recommended traversable area segmentation using automatically labeled images for autonomous driving in pedestrian environment with no edges. *Sensors (Switzerland)* 21 (2), 1–22. <http://dx.doi.org/10.3390/s21020437>.
- Osman, H., Darwish, N., Bayoumi, A., 2023. PlaceNet: A multi-scale semantic-aware model for visual loop closure detection. *Eng. Appl. Artif. Intell.* 119, 105797.
- Papadakis, P., 2013. Terrain traversability analysis methods for unmanned ground vehicles: A survey. *Eng. Appl. Artif. Intell.* 26 (4), 1373–1385.
- Pohlen, T., Hermans, A., Mathias, M., Leibe, B., 2017. Full-resolution residual networks for semantic segmentation in street scenes. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp.* 4151–4160.
- Romera, E., Alvarez, J.M., Bergasa, L.M., Arroyo, R., 2017. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Trans. Intell. Transp. Syst.* 19 (1), 263–272.
- Romero, A., Rister, S., 2017. ERFNet: Efficient residual factorized network for real-time semantic segmentation. In: *Proceedings of the IEEE International Conference on Computer Vision. ICCV*, URL <https://arxiv.org/abs/1703.06895>.
- Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18. Springer*, pp. 234–241.
- Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M., 2016. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp.* 3234–3243.
- Rothrock, B., Papon, J., Kennedy, R., Ono, M., Heverly, M., Cunningham, C., 2016. SPOC: Deep learning-based terrain classification for Mars rover missions. *AIAA Space Astronaut. Forum Expo. SPACE 2016 (September)*, <http://dx.doi.org/10.2514/6.2016-5539>.
- Sancho-Pradel, D.L., Gao, Y., 2010. A survey on terrain assessment techniques for autonomous operation of planetary robots. *JBIS-Journal Interplanet. Soc.* 63 (5–6), 206–217.
- Sharma, S., Dabbur, L., Hannis, T., Mason, G., Carruth, D.W., Doude, M., Goodin, C., Hudson, C., Ozier, S., Ball, J.E., et al., 2022. Cat: Caves traversability dataset for off-road autonomous driving. *IEEE Access* 10, 24759–24768.
- Shojaiee, F., Baleghi, Y., 2023. EFASPP U-net for semantic segmentation of night traffic scenes using fusion of visible and thermal images. *Eng. Appl. Artif. Intell.* 117, 105627.
- Shu, Y., Dong, L., Liu, J., Liu, C., Wei, W., 2024. Overview of terrain traversability evaluation for autonomous robots. *J. Field Robot.*
- Song, S., Huang, T., Li, C., Shao, G., Gao, Y., Zhu, Q., 2023. A safety-assured semantic map for an unstructured terrain environment towards autonomous engineering vehicles. *Drones* 7 (9), 550.
- Song, S., Lichtenberg, S.P., Xiao, J., 2015. Sun rgb-d: A rgb-d scene understanding benchmark suite. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp.* 567–576.
- Takikawa, T., Acuna, D., Jampani, V., Fidler, S., 2019. Gated-scnn: Gated shape cnns for semantic segmentation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision. pp.* 5229–5238.
- Valada, A., Mohan, R., Burgard, W., 2020. Self-supervised model adaptation for multimodal semantic segmentation. *Int. J. Comput. Vis.* 128 (5), 1239–1285.
- Valada, A., Oliveira, G.L., Brox, T., Burgard, W., 2017. Deep multispectral semantic scene understanding of forested environments using multimodal fusion. In: *2016 International Symposium on Experimental Robotics. Springer*, pp. 465–477.
- Wang, H., Ma, B., Yu, L., Chen, M., Wang, L., Li, H., 2025. Traversability analysis for tracked vehicles in unstructured environments: An approach employing vehicle traversing capability and terrain's multi-modal data. *Automot. Innov.* 8, 264–280. <http://dx.doi.org/10.1007/s42154-024-00340-w>.
- Wellhausen, L., Dosovitskiy, A., Ranftl, R., Walas, K., Cadena, C., Hutter, M., 2019. Where should i walk (Predicting terrain properties from images via self-supervised learning. *IEEE Robot. Autom. Lett.* 4 (2), 1509–1516. <http://dx.doi.org/10.1109/LRA.2019.2895390>.
- Wolf, P., Vierling, A., Ropertz, T., Berns, K., 2020. Advanced scene aware navigation for the heavy duty off-road vehicle unimog. In: *IOP Conference Series: Materials Science and Engineering*, vol. 997, (1), IOP Publishing, 012093.
- Wu, J., Kuang, H., Lu, Q., Lin, Z., Shi, Q., Liu, X., Zhu, X., 2022. M-FasterSeg: An efficient semantic segmentation network based on neural architecture search. *Eng. Appl. Artif. Intell.* 113, 104962.
- Wu, X., Xu, N., Zhou, J., 2000. Terrain Classification Algorithm for Off-Road Vehicle Using 1D CNN-LSTM Network. *Unkn. Journal/Conference*.
- Yang, M., Yu, K., Zhang, C., Li, Z., Yang, K., 2018. Denseaspp for semantic segmentation in street scenes. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp.* 3684–3692.
- Yu, F., Xie, Y., 2018. BiSeNet: Bilateral segmentation network for real-time semantic segmentation. In: *Proceedings of the European Conference on Computer Vision. ECCV*, URL <https://arxiv.org/abs/1808.08489>.
- Yu, H., Yang, S., Gu, W., Zhang, S., 2017. Baidu driving dataset and end-To-end reactive control model. *IEEE Intell. Veh. Symp. Proc. (Iv)*, 341–346. <http://dx.doi.org/10.1109/IVS.2017.7995742>.
- Zhang, M., Ma, L., Wu, Y., Shen, K., Sun, Y., Leung, H., 2025. High-traversability and precise navigation for mobile robots in constrained environments. *IEEE Sensors J.* <http://dx.doi.org/10.1109/JSEN.2025.3566035>.
- Zhang, X., Qiao, Y., 2017. MobileUNet: Efficient semantic segmentation for mobile devices. In: *Proceedings of the IEEE International Conference on Computer Vision. ICCV*, URL <https://arxiv.org/abs/1705.02148>.
- Zhao, H., Shi, J., Qi, X., 2017. Pyramid scene parsing network. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR*, pp. 6230–6239, URL <https://arxiv.org/abs/1704.03816>.



DSPACE

<https://dspace.org/>

On real-time semantic segmentation with comprehensive off-road datasets for enhanced terrain classification

Beycimen, Semih; Art, Ali; Unal, Mehmet; Ignatyev, Dmitry; Zolotas, Argyrios

2026-02-15

Elsevier

<https://doi.org/10.1016/j.engappai.2025.113680>

Beycimen S, Art A, Unal M, et al., (2026) On real-time semantic segmentation with comprehensive off-road datasets for enhanced terrain classification. Engineering Applications of Artificial Intelligence, Volume 166, Part B, February 2026, Article number 113680