2020 NDIA GROUND VEHICLE SYSTEMS ENGINEERING AND TECHNOLOGY SYMPOSIUM
<AUTONOMY, ARTIFICIAL INTELLIGENCE & ROBOTICS (AAIR)> TECHNICAL SESSION
AUGUST 11-13, 2020 - NOVI, MICHIGAN

SEMANTIC DIGITAL SURFACE MAP TOWARDS COLLABORATIVE OFF-ROAD VEHICLE AUTONOMY

Howard J. J. Brand, Bing Li

Department of Automotive Engineering, Clemson University International Center for Automotive Research (CU-ICAR), Greenville, SC 29607, USA

ABSTRACT

The fundamental aspect of unmanned ground vehicle (UGV) navigation, especially over off-road environments, are representations of terrain describing geometry, types, and traversability. One of the typical representations of the environment is digital surface models (DSMs) which efficiently encode geometric information. In this research, we propose a collaborative approach for UGV navigation through unmanned aerial vehicle (UAV) mapping to create semantic DSMs, by leveraging the UAV wide field of view and nadir perspective for map surveying. Semantic segmentation models for terrain recognition are affected by sensing modality as well as dataset availability. We explored and developed semantic segmentation deep convolutional neural networks (CNN) models to construct semantic DSMs. We further conducted a thorough quantitative and qualitative analysis regarding image modalities (between RGB, RGB+DSM and RG+DSM) and dataset availability effects on the performance of segmentation CNN models.

Citation: H. J. J. Brand, B. Li, "Semantic Digital Surface Map Towards Collaborative Off-Road Vehicle Autonomy", In *Proceedings of the Ground Vehicle Systems Engineering and Technology Symposium* (GVSETS), NDIA, Novi, MI, Aug. 11-13, 2020.

1. INTRODUCTION

Important to path planning and navigation for off-road ground vehicle systems are digital surface models (DSM). Given a 2D grid space perpendicular to the Earth's surface, DSMs report the elevation of the ground and objects on ground for each grid point. This encodes 3D geometric information about the terrain as well as size/shape and location of objects and obstacles. 3D terrain information is useful for determining traversability. Braun et al [1] used elevation models and slope information to describe the traversability of urban environments. They validated their method in simulation and field measurements. Ohki et al [2] developed a path planning method to allow a UGV to navigate rough volcanic terrain using extended elevation models. Guastella et al [3] also used DSMs to conduct traversability estimation in urban environments and performed global path planning for disaster response scenarios.

DSMs also provide information that allow efficiency objectives to be achieved in path planning. Hameed et al [4] developed a path

planning method based on 3D coverage efficiency from DSM information. This allows for more efficient coverage and reduced skipping and overlap during robotic harvesting procedures. Spekken et al [5] developed a method that utilized elevation information to reduce disturbances to the environment, such as soil erosion, during navigation. DSMs are also useful for encoding object geometry and location. Oniga et al [6] used local elevation maps and object detection from a ground vehicle to develop surface models of objects in order to perform path planning and obstacle avoidance.

Unmanned aerial vehicles (UAVs) have a larger view of land cover and are generally used to develop large-scale DSMs which are crucial towards implementing global path planning for unmanned ground vehicles (UGVs) [3], [7], [8]. UAVs have less limitations in mobility and benefit from a wider viewing perspective than UGVs. However, UGV's payload capacity allows them to be better candidates for interacting with the environment and achieving missions [7]. For outdoor, urban environments UGVs must navigate through prepared and unprepared terrain. Many of the existing traversability methods for urban scenarios that use DSMs are usually restricted to implementation on homogeneous terrain. To take advantage of existing DSM-based UGV navigation methods to on-road and off-road terrain scenarios a description of the terrain and corresponding obstacles common to both terrain environments are needed as well as the DSM information. This can be provided by a semantic DSM with terrain types and corresponding terrain obstacle labels.

A number of works have been conducted on using machine learning-based approaches to develop semantic land cover segmentation. While this work has been developed to explore the effectiveness of different strategies to obtain segmentation results, there hasn't been a focus regarding performance generalization in relation to data availability and applicability to unexplored environments. In our work we explore the use of a UAV equipped with an RGB (red, green, and blue channel) camera and a Lidar sensor to map the Clemson University International Center for Automotive Research campus (CU-ICAR). Our work used a land cover semantic segmentation network for three available sensing modalities in our platform RGB, RGBDSM, RGDSM, and RGDSM+ (representing an extended dataset for the RGDSM sensing modality).

2. BACKGROUND

There have been a lot of efforts payed in recent years to develop semantic maps from aerial observations, using multiple sensing modalities. This has led to the development of many classical machine learning and deep learning approaches. Salih et al [9] trained a Maximum Likelihood classifier on principal components of the of satellite images of the Al-Ahsaa Oasis to classify bare soil, sand, urban, vegetation, and water. The satellites images were composed of six bands including near-infrared (NIR), red, green, and blue spectra. Feng et al [10] used Random Forests (RF) on RGB images and texture feature maps to classify bare soil, grass, trees, shrubs, water, and impervious surfaces. Liu et al [11] used conditional random fields to combine multi-view information and context to improve the accuracy of the semantic segmentation classifiers from RGB data. The study tested RF, Gaussian Mixture Model (GMM), Support Vector Machine (SVM), and Deep Convolutional Neural Network (DCNN) classifiers. RF and DCNN classifiers were found to achieve the greatest accuracy.

Šćepanovic et al [12] utilized DCNNs for semantic segmentation of C-band synthetic aperature radar (SAR) images. They tested the U-Net, SegNet, DeepLabV3+, BiSeNet, FRRN-B, FC-DenseNet, and PSPnet. Al-Najjar [13] used a DCNN to segment fused RGB plus DSM (digital surface elevation) images.

With the various kinds of land cover image modalities and datasets, many works seek to investigate the contributions of certain sensor modes to the classification task. This is important as incorporating unrelated inputs to the statistical learning tasks degrades classifier performance. Much of these considerations are explored in classical machine learning approaches. Salih et. al [9] reported the eigenvalues (variances) or contributions of each channel to the principal component inputs to the Maximum Likelihood classifier. Feng et. al [10] was able to determine the variable importance through a perturbation test of the inputs to the RF classifier. This directly allowed contribution to classifier performance of each input to be analyzed.

With DCNNs this is a more challenging question as there are a very large amount features influencing the classification outcome. Additionally, these features are learned parameters where their form or structure are largely unknown and black box. In the work of Salih et al [9] the performance of a segmentation CNN with RGB input was compared to that of an RGB plus DSM network. It was found that the RGB plus DSM input-based CNN outperformed the RGB input based CNN.

According to classical machine learning methods the B channel is one of the most crucial variables for landcover classification when compared to texture-based information and multiple spectral channels such as R, G, and near-infrared (NIR) [9], [10]. The availability of datasets is also crucial to generalized segmentation performance. In our study we investigate the contribution of inputs vs the contribution of data for DNN based classification. We further studied this trade-off by comparing the performances of four different sensing modalities: RGB, RGBDSM, RGDSM and RGDSM+ based networks. Effects of pretraining were also incorporated in the analysis.

3. METHODOLOGY

This work deals with two fundamental problems in statistical learning: covariate importance and the availability of data. For our approach we consider platforms with high resolution RGB camera and

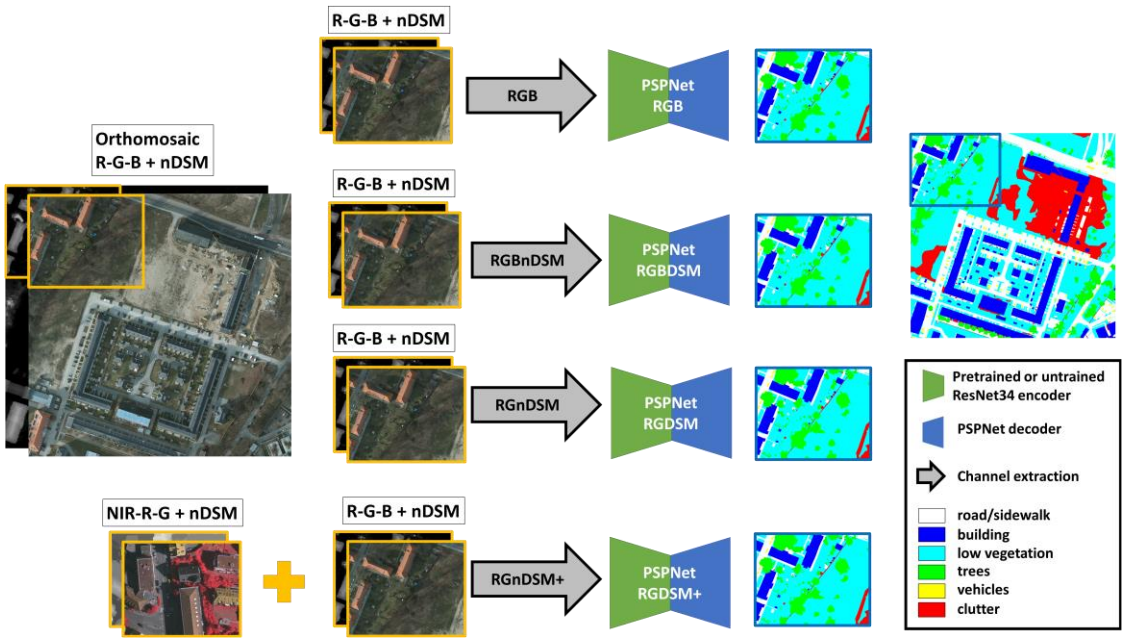

Figure 1: Multiple Image Mode PSPNet Setup.

LIDAR sensing modalities, where we want to semantically label the data according to the International Society for Photogrammetry and Remote Sensing (ISPRS) [14] classes: low vegetation, trees, roads, buildings, vehicles, and clutter. There are two ISPRS urban environment datasets that support this labeling application, the ISPRS 2D and ISPRS 3D Semantic Labeling datasets. The 2D dataset was acquired with three main image mode sets: NIR-R-G, RGB, and DSM. There were two main image modes in the 3D dataset: NIR-R-G and DSM. In relation to the UAV platform there are two main modalities of interest in relation to the available dataset. There is one dataset for the full set of covariates R, G, B and DSM. There are two datasets for a subset of covariates R, G, and DSM.

With this study we are using PSPNet [15] with a ResNet34 encoder for both pretrained (on ImageNet [16]) and untrained cases. The input to the network is a 512x512xC, where C represents the number of input channels depending on the image modality. C is governed by three sensors modes, RGB, RGBnDSM, and RGnDSM. The channel nDSM represents the DSM data normalized by regional minimum elevations to have a minimum elevation of zero. Class predictions made on orthographic images are conducted by making predictions on 512x512 sections. An illustration of the network is shown in Figure 1.

The 2D Semantic Dataset was acquired in Potsdam, Germany during Autumn/Winter months with many examples of leafless trees. This dataset is composed of 6000x6000 orthomosaic images. For this study eleven orthomosaic images are used to generate the training andvalidation datasets for the RGB, RGBnDSM, and RGnDSM sensor modes. The training/validation images are generated by randomly sampling 512x512xC images from the larger orthomosaic images. The training and validation dataset consisted of 14,000 and 9,000 images respectively.

Additional RGnDSM training/validation data were developed from the 3D Semantic Dataset. The 3D Semantic Dataset consists of 16 orthographic images and was acquired in Vaihingen, Germany during the Spring/Summer months, containing examples of full, leafy trees. The combined larger RGnDSM training and validation dataset (termed RGnDSM+) consisted of 33,000 and 20,000 images respectively. Finally, eleven other orthographic images from the 2D Semantic Dataset, not used to generate the training/validation set, were used to test the network performance.

Regarding the in-field data, a UAV instrumented with an RGB camera and LIDAR sensor flew a flight path over the Clemson University International Center of Automotive Research (CU-ICAR) campus. This campus is a testing site for self-driving vehicles for both on-road and off-road environments. RGB, RGBnDSM, and RGnDSM orthographic images were developed from the data to allow for testing the generalization performance.

4. EXPERIMENTS AND RESULTS

The class predictions from the PSPNet segmentation network are according to the International Society for Photogrammetry and Remote Sensing (ISPRS) [14] class labels and color codes: low vegetation (cyan, essentially off-road), trees (green), impervious surfaces (white), buildings (blue), vehicles (yellow), and clutter (red).

Four PSPNets were trained on four datasets, RGB, RGBnDSM, RGnDSM, and RGnDSM+. The classification and segmentation performance of the networks were evaluated using the mean intersection-over-union (mIOU) criteria:

$$mIOU(C) = \frac{1}{n}\sum_{i=0}^{n}\frac{pixel\_sum\left(L_G^C \cap L_{PSP}\left(X_i^C\right)\right)}{pixel\_sum\left(L_G^C \cup L_{PSP}\left(X_i^C\right)\right)} \quad (1)$$

Where L_G^C represents the ground truth class image and the $L_{PSP}(\cdot)$ represents the network class image. According to this metric an ideal segmentation map would produce a mIOU value of 1 while an incorrect segmentation would result in a mIOU value tending towards 0.

4.1. Semantic Segmentation Performance

Table 1 shows the mIOU per class for the RGB, RGBnDSM, RGnDSM, and RGnDSM+ image modality networks using a pretrained ResNet34 encoder. Inspection of Table 1 reveals that the RGBnDSM, RGnDSM, and RGnDSM+ networks outperform the RGB network with overall classification score of the 0.692, 0.707, and 0.685 for the RGBnDSM, RGBnDSM, and RGnDSM+ modalities respectively. Further inspection reveals that the DSM channel contributes to increasing accuracy for building and tree classes.

Table 2 shows the mIOU per class for the RGB, RGBnDSM, RGnDSM, and RGnDSM+ image modality networks using a ResNet34 encoder without pretraining. Inspection of Table 2 shows similar contributions associated with involving DSM information with increases in building and tree classification accuracy.

A comparison between Table 1 and Table 2 reveals identical classification performance for the testing dataset for all sensing modalities. Based on the testing set results pretraining offers no distinguishable effect to generalized classification performance.

The RGnDSM modality has the highest classification performance according to the testing dataset results. A comparison between the RGnDSM and RGnDSM+ networks shows a reduction in classification performance accuracy in the RGnDSM+ network for every class except the vehicle class. This relates to the test set being from Potsdam and the RGnDSM+ dataset incorporating images from Vaihingen, Germany. This points to a challenge related to evaluating the generalization of the landcover classification performance.

Because of the considerable resources necessary to develop landcover data, it is generally common practice to develop datasets for single regions or cities and incorporate a subset of the dataset to be the evaluation set. This however has implications on the generalization of the testing data. This is especially relevant to the UGV navigation chain when applying these networks to provide a semantic DSM of an unknown environment.

Table 1: mIOU Class Performance. With pretrained ResNet34 encoder

	Buildings	Clutter	Trees	Low Vegetation	Vehicles	Roads/ Sidewalk	Overall
RGB	0.864	0.275	0.668	0.682	0.811	0.778	0.680
RGnDSM	0.906	0.361	0.703	0.678	0.807	0.785	0.707
RGBnDSM	0.909	0.284	0.683	0.683	0.808	0.786	0.692
RGnDSM+	0.897	0.271	0.691	0.662	0.811	0.779	0.685

Table 2: mIOU Class Performance. Without pretrained ResNet34 encoder

	Buildings	Clutter	Trees	Low Vegetation	Vehicles	Roads/ Sidewalk	Overall
RGB	0.861	0.283	0.670	0.687	0.804	0.779	0.681
RGnDSM	0.900	0.363	0.706	0.685	0.801	0.785	0.707
RGBnDSM	0.907	0.292	0.684	0.687	0.805	0.788	0.694
RGnDSM+	0.897	0.278	0.693	0.668	0.803	0.780	0.686

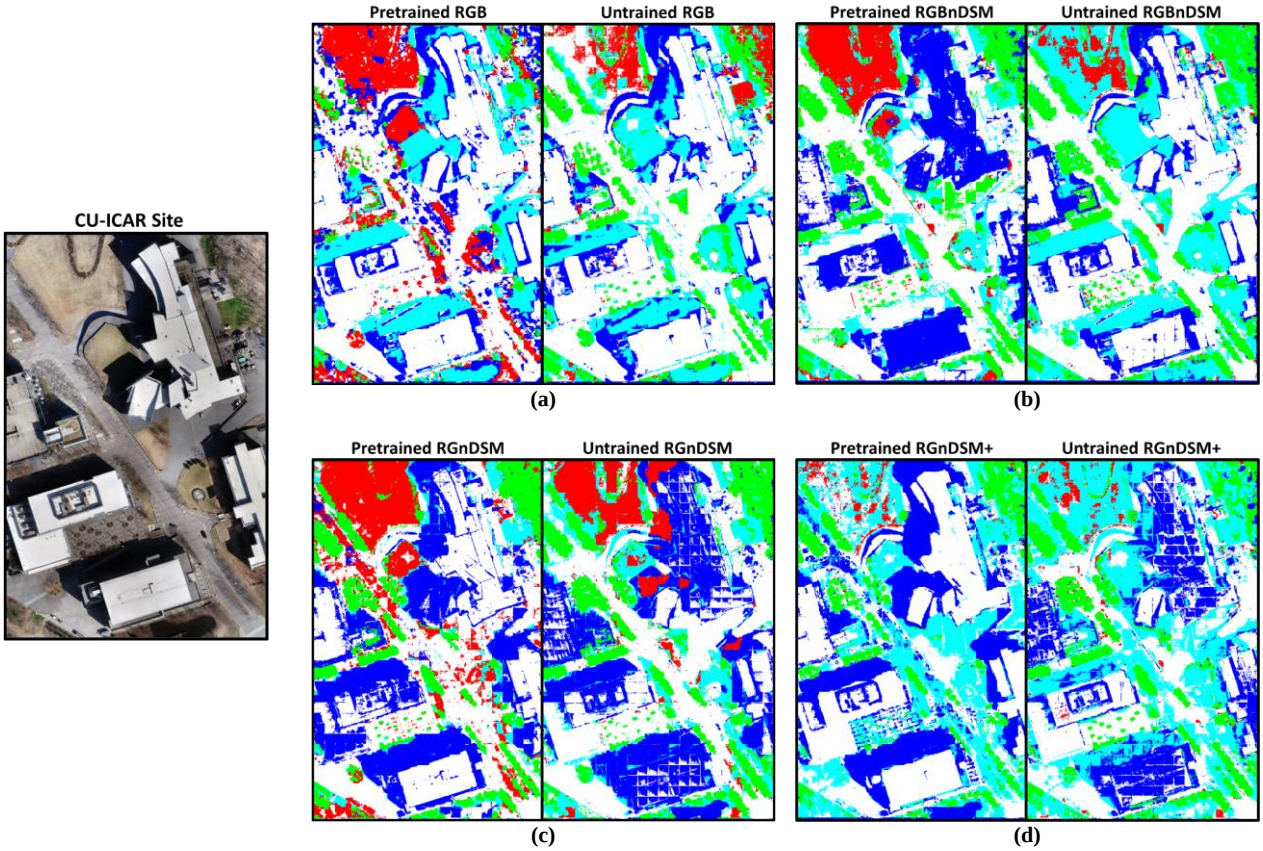

Figure 2: ICAR Segmentation Results. Low vegetation (cyan, essentially off-road), trees (green), impervious surfaces (white), buildings (blue), vehicles (yellow), and clutter (red).

4.2. In-Field Data Test of CU-ICAR Campus

Both the pretrained and untrained counter parts of the RGB, RBGnDSM, RGnDSM, and RGnDSM+ networks were used to segment the dataset of Clemson University International Center for Automotive Research (CU-ICAR) site. The CU-ICAR campus was developed based on innovative architectural designs which provide a number of challenging artifacts not represented in the training datasets. The road leading in the campus is made of multi-colored brick instead of asphalt concrete. The sidewalks are also made of the same material. Many of the buildings have different artistic vented metal canopies outlining the roofs and many roofs have multiple cascading levels.

Figures 2a-2d are comparative illustrations between pretrained and untrained RGB, RGBnDSM, RGnDSM, and RGnDSM+ networks respectively.

The segmentation performance differences between the pretrained networks and the non-pretrained networks are very apparent despite the similarities between Tables 1 and 2. There are very few and very small-scale true instances of clutter in the ICAR dataset except a slender winding creek near the top of the image. Between the pretrained and untrained networks there are far fewer misclassifications of clutter in the untrained networks for each sensing modality.

Each network has difficulty distinguishing shadow from buildings with the RGBnDSM cases and the untrained RGnDSM+ case having the most success. It is also difficult for the networks to classify the tops of ICAR buildings correctly. They are mostly misclassified as roads by all the networks with the exception of the pretrained RGBnDSM and the untrained RGnDSM+ having the least difficulty.

From these results, the full RGBnDSM input network was able to overcome some classification challenges with pretraining however at the expense of instability in the clutter class. This instability may be attributed to the parameters being pretrained on an image classification task rather than a segmentation task. While the RGDSM+ network was better at detecting low vegetation than the RGDSM network it was not able to detect roads as well. Untrained versions of the RGDSM and RGDSM+ networks performed better than their pretrained counterparts.

5. CONCLUSIONS

An analysis was conducted regarding the generalization of landcover classification networks and their applicability to providing semantic DSMs from UAV observations for global UGV navigation in unknown environments. The problems studied in this work were related to the availability of semantic datasets for specific class labels and sensor modalities. This creates a tradeoff between training DNN on a limited dataset composed of a full set of inputs available on the UAV platform and training on a larger dataset composed of a subset of inputs available. The effects of pretraining on generalized performance were also explored along with common methods for determining network performance generalization.

Based on the results of the study it was determined that common methods for determining the generalization of the network performance most likely should include dataset from unobserved cities or entirely different regions. Testing datasets created from a subset of city-wide or region-wide datasets are closely related to the training and validation datasets and are not good predictors of generalized performance.

It was also observed that pretraining can potentially aid in improving segmentation performance and overcome data unavailability at the expense of prediction instability. Performance enhancements to segmenting a subset of the available input sensor modality on additional datasets depends on the nature of the additional dataset and whether the missing inputs are important observers for classification performance. For future work, this tradeoff principle will be evaluated on another dataset containing all or a subset of the desired labels.

REFERENCES

[1] T. Braun, H. Bitsch, and K. Berns, “Visual terrain traversability estimation using a combined slope/elevation model,” in *Annual Conference on Artificial Intelligence*, 2008, pp. 177–184.

[2] T. Ohki, K. Nagatani, and K. Yoshida, “Path planning for mobile robot on rough terrain based on sparse transition cost propagation in extended elevation maps,” in *2013 IEEE International Conference on Mechatronics and Automation*, 2013, pp. 494–499.

[3] D. C. Guastella, L. Cantelli, C. D. Melita, and G. Muscato, “A Global Path Planning Strategy for a UGV from Aerial Elevation Maps for Disaster Response.,” in *ICAART (1)*, 2017, pp. 335–342.

[4] I. A. Hameed, A. la Cour-Harbo, and O. L. Osen, “Side-to-side 3D coverage path planning approach for agricultural robots to minimize skip/overlap areas between swaths,” *Rob. Auton. Syst.*, vol. 76, pp. 36–45, 2016.

[5] M. Spekken, S. De Bruin, J. P. Molin, and G.

Sparovek, “Planning machine paths and row crop patterns on steep surfaces to minimize soil erosion,” *Comput. Electron. Agric.*, vol. 124, pp. 194–210, 2016.

[6] F. Oniga and S. Nedevschi, “Processing dense stereo data using elevation maps: Road surface, traffic isle, and obstacle detection,” *IEEE Trans. Veh. Technol.*, vol. 59, no. 3, pp. 1172–1182, 2009.

[7] T. H. Nam, J. H. Shim, and Y. I. Cho, “A 2.5 D map-based mobile robot localization via cooperation of aerial and ground robots,” *Sensors*, vol. 17, no. 12, p. 2730, 2017.

[8] A. M. Khaleghi, D. Xu, S. Minaeian, Y. Yuan, J. Liu, and Y.-J. Son, “Analysis of uav/ugv control strategies in a dddams-based surveillance system,” in *IIE Annual Conference. Proceedings*, 2015, p. 2283.

[9] A. Salih, “Classification and mapping of land cover types and attributes in Al-Ahsaa Oasis, Eastern Region, Saudi Arabia Using Landsat-7 Data,” *J. Remote Sens. GIS*, vol. 7, p. 228, 2018.

[10] Q. Feng, J. Liu, and J. Gong, “UAV remote sensing for urban vegetation mapping using random forest and texture analysis,” *Remote Sens.*, vol. 7, no. 1, pp. 1074–1094, 2015.

[11] T. Liu, A. Abd-Elrahman, A. Zare, B. A. Dewitt, L. Flory, and S. E. Smith, “A fully learnable context-driven object-based model for mapping land cover using multi-view data from unmanned aircraft systems,” *Remote Sens. Environ.*, vol. 216, pp. 328–344, 2018.

[12] S. Šćepanović, O. Antropov, P. Laurila, V. Ignatenko, and J. Praks, “Wide-Area Land Cover Mapping with Sentinel-1 Imagery using Deep Learning Semantic Segmentation Models,” *arXiv Prepr. arXiv1912.05067*, 2019.

[13] H. A. H. Al-Najjar *et al.*, “Land cover classification from fused DSM and UAV images using convolutional neural networks,” *Remote Sens.*, vol. 11, no. 12, Jun. 2019.

[14] “The ISPRS data set collection.” .

[15] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*, 2009, pp. 248–255.