Forest Terrain Identification using Semantic Segmentation on UAV Images

Muhammad Umar [*1] **Lakshmi Babu Saheer** [*1] **Javad Zarrin** [1]

Abstract

Beavers' habitat is known to alter the terrain, providing biodiversity in the area, and recently their lifestyle is linked to climatic changes by reducing greenhouse gases levels in the region. To analyse the impact of beavers' habitat on the region, it is, therefore, necessary to estimate the terrain alterations caused by beaver actions. Furthermore, such terrain analysis can also play an important role in domains like wildlife ecology, deforestation, land-cover estimations, and geological mapping. Deep learning models are known to provide better estimates on automatic feature identification and classification of a terrain. However, such models require significant training data. Pre-existing terrain datasets (both real and synthetic) like CityScapes, PASCAL, UAVID, etc, are mostly concentrated on urban areas and include roads, pathways, buildings, etc. Such datasets, therefore, are unsuitable for forest terrain analysis. This paper contributes, by providing a finely labelled novel dataset of forest imagery around beavers' habitat, captured from a high-resolution camera on an aerial drone. The dataset consists of 100 such images labelled and classified based on 9 different classes. Furthermore, a baseline is established on this dataset using state-of-the-art semantic segmentation models based on performance metrics including Intersection Over Union (IoU), Overall Accuracy (OA), and F1 score.

1. Introduction

Beavers are well-known for changing the existing ecosystem and terrain by cutting down trees, digging canals, constructing dams and lodges on streams, thereby, creating wetlands and ponds. Recent studies (Nummi et al., 2018) have indicated that building dams increase the water levels in the area and as a result, reduces the carbon levels by absorbing carbon directly from soil and carbon dioxide from the air. This carbon is further dissolved in the soil to be later used by plants or is transferred downstream. To estimate the impact on climate, it is, therefore, necessary to first identify the changes in the terrain around beavers' habitat caused by beaver actions. Hence, the main aim of this research is to analyse the performance of pre-existing deep learning models to perform classification and recognition of local textural patterns like shrubs, beaver lodges and dams, vegetation, trees, etc within terrain images.

Legacy computer vision methods like colour histogram and estimating the colour frequency and other feature map extraction models like Random Forests (RF) (Breiman, 2001) and Conditional Random Fields (CRFs) (Lafferty et al., 2001) etc were used previously however such algorithms are mostly ineffective as terrain images are susceptible to climatic and locality-based changes. Furthermore, colour characteristics within the same class make them indistinguishable from each other, sometimes, even for a human eye.

Semantic segmentation assigns each pixel in an image a specific class label, which is the core requirement of our classification problem. Semantic segmentation divides the data in the domain into smaller units like superpixels, super voxels, grid-based units, etc. Object detection, in contrast, uses a template matching algorithm and creates a bounding box over units based on the correlation between the matching template and the pixel data. Such a bounding box never tightly fits the detected classes and hence not usable on terrain patterns captured by aerial drone imagery, where the size of such objects is extremely small. Deep neural networks outperform any other frameworks used in computer vision for solving problems in domains like pattern recognition, feature extraction, and detection/classification. Because of the complexity of the patterns in terrain, current classification problem requires proper extraction of features from images for classification. One of the most important challenges thus, using such models, is the lack of datasets to train deep models. The key contributions of this paper, thus, include **(a)** A novel dataset that consists of 100 high-resolution images taken from UAV near beaver habitat. The

[*]Equal contribution [1]Anglia Ruskin University, Cambridge, United Kingdom. Correspondence to: Muhammad Umar <mu283@student.aru.ac.uk>, Lakshmi Babu Saheer <lakshmi.babu-saheer@aru.ac.uk>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

images are labelled for 9 distinct classes. Images are split into two parts, training and validation. Training contains 70 images whereas validation and testing contain 30 images. Each image is 18 MB with 300 dpi and pixels size is 5472*3648. **(b)** A benchmark on the proposed dataset by evaluating multiple state-of-the-art segmentation models. The evaluation is performed using 3 different metrics namely, Overall accuracy (OA), Dice coefficient (F1 Score), and Jaccard index (Intersection over union - IOU).

The rest of this paper organized as follows: Section 2 provides an overview of the current literature, Section 3 introduces our proposed dataset for forest terrain identification containing patterns that contribute to the climatic conditions, Section 4 discusses our evaluation approach and a benchmark for the proposed dataset, Section 5 presents evaluation results and a discussion of results, and finally, Section 6 concludes the paper and discusses future works.

2. Literature Review

FCN (Shelhamer et al., 2016) performed segmentation tasks with high accuracy by performing an "end-to-end" training mechanism. Such convolutional networks take any image of any input size based on the model parameters and then output the same image with segmented masks applied. (Zürn et al., 2020) explored the self-supervised learning in the domain of terrain identification for self-autonomous robots. (Kattenborn et al., 2021) used different CNN-based architectures to identify vegetation via remote sensing.

A powerful deep network U-Net (Ronneberger et al., 2015) introduced skip connections and residual networks to achieve high accuracy on a biomedical dataset. The model can be enhanced using multiple architectural backbones and weights from different models. SegNet (Badrinarayanan et al., 2016) was an upgrade to UNet. Instead of passing complete features to the next layer, only the max pooled version of features was passed on thereby increasing performance. Similarly, basic computer-vision-related tasks like rotation equivariance were performed to segment high-resolution images captured from a direct flight path. (Arun et al., 2019) created dataset using super-high-resolution images from drones to train different CNN however the dataset mostly consists of mild areas and is not suitable for forest terrains. (Fikri et al., 2019) used CNN to cluster trees as superpixels and then perform pixel-based segmentation using colour threshold.

Simple Linear Iterative Clustering (SLIC) (Achanta et al., 2012), and Simple Non- Iterative Clustering (SNIC) (Achanta & Susstrunk, 2017) have successfully been implemented to generate powerful superpixel based segmentations. SLIC performs k-means on CIELab colour code instead of using RGB images to generate macro superpixels. The SNIC algorithm instead of using k-means like SLIC clusters pixels by explicitly enforcing connectivity from the start.

Since superpixels make a macro pixel of localized area, adding CNN to train such models reduces the complexity and improves efficiency for semantic segmentation. Such a model was used by (Yang et al., 2020) where superpixels were used with CNN on a video stream of 50fps. Superpixels were then down-up sampled using autoencoder approach to generate predictive masks. (Wang et al., 2019) used similar approach however they introduced two datasets to classify porifera region in water. (Chen et al., 2019) used multiple superpixel methods to classify land cover area using deep neural networks. However, the problem remains challenging.

3. Proposed Dataset

The goal of this research is to perform a quantitative analysis of different terrain patterns for climate analysis using state-of-art models. To this end, the dataset must contain all possible patterns that contribute to the climatic conditions. It is also important to capture and identify data that must not be biased to a single class as some patterns can be in excessive quantity in a terrain than others. For forest terrain identification, no such dataset so far, to our best of knowledge, is available. The terrain identification requires images, to be captured from an aerial view to encompass a large area. Some of the already existing datasets available are CityScapes (Cordts et al., 2016) PASCAL (Everingham et al., 2015) UAVID (Lyu et al., 2020) etc. However, none of the above datasets contain images related to forest imagery and are mostly concentrated on urban regions, containing buildings, road pathways, etc.

3.1. Data Collection

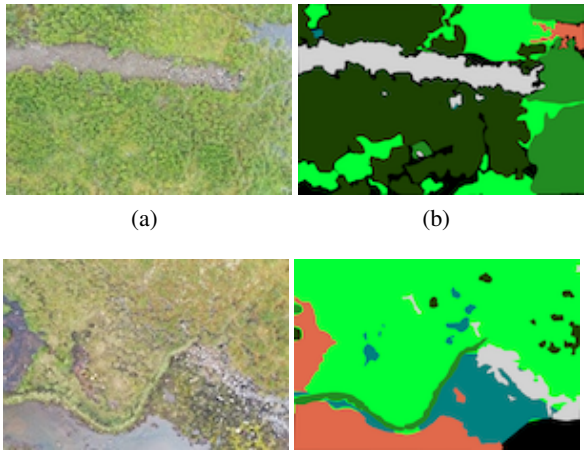
To collect data, UAVs are used with mounted cameras having a resolution of 18 megapixels. A single image resolution is 5472*3648 with each image size being approximately 13 MB. The images are captured on a medium sunny day to acquire maximum details of the terrain. For annotations, multiple 3rd party tools are available. Some of the famous tools are QGIS (QGIS, 2020), ArcGIS (ArcGis, 2020) and LabelBox (LabelBox, 2020). After careful consideration and experiments, we found LabelBox to be easier to use. The annotations are performed on 100 different images with all images selected in a manner to create a balanced dataset for all classes.

The terrain is classified into 10 different classes and their percentage is dataset is shown in Table 1. ROI (Region of interest) annotations were performed on images on maximum zoom to provide the best possible results and therefore took

(a) Stones (b) Water (c) Grass (d) Dam

(e) Shrub (f) Large Shrubs (g) Vegetation (h) Soil

Figure 1. Dataset classes



(a) (b) (c) (d)

Figure 2. (a) and (c) are sample images. (b) and (d) are corresponding generated masks.

nearly 30 min - 1 hour on an image based on the complexity of terrain in each image. Sample images and their corresponding generated masks are shown in Fig 2 and Fig 2. Each image was labelled in maximum resolution; therefore each image is subdivided into 240 parts keeping aspect ratio same to avoid any distortions. Each image thus generated has a resolution of 342*243. Furthermore, these images are subjected to cropping to adjust their sizes to 340*240. This action is performed to avoid size differences, caused by applying pooling layers in deep models.

3.2. Dataset Classes

Table 1. shows the chosen class attributes and their percentage of occurrences in the dataset. It should be noted, even if the images were high resolution, there are some instances where small terrain patches are not clearly recognized and

Table 1. Distribution of classes in the dataset

CLASS	TRAIN(%)	VALIDATION(%)
UNKNOWN	1.58	1.64
SHRUBS	15.72	13.83
GRASS	14.05	13.74
LARGE SHRUBS	10.22	11.22
VEGETATION	13.41	13.16
STONES	11.38	10.43
WATER	12.35	13.73
SOIL	13.05	14.26
BEAVER DAM	8.24	7.99

are marked with class `unknown`. The datasets are marked in a fashion to keep such annotations minimum. Furthermore, corresponding class masks are given zero RGB values to avoid any conflicts with other classes. Sample class images are shown in Fig 2.

4. Evaluations

Evaluation of the performance of any deep learning model on a dataset poses different challenges. The models may provide good results depending on measured performance metrics. However, the real performance (e.g. accuracy) can be quite different when wrong metrics are selected, therefore, it is necessary to test models on multiple metrics. To evaluate the state-of-art models on our dataset, the models are evaluated using 3 different metrics and the result are summarized in a table.

4.1. Performance Metrics

Pixel accuracy is a quantitative metric that calculates the percentage of correct pixels classified for a class compared with the actual ground truth. This metric is greatly affected by `class imbalance`. The metric assigns equal weight

Table 2. Pre-trained Models used for Evaluation

Model	Description
UNet (Ronneberger et al., 2015)	An Encoder-Decoder architecture with Skip-Connections. The encoder is feed-forward CNN (backbone) to extract feature maps of images. Decoder upsamples the feature map to generate back the information.
FPN (Lin et al., 2017)	Bottom-up pathway is a CNN (backbone) to extract feature maps. Using pyramid modelling strategy, low-resolution feature map from higher layers are merged with high-resolution feature map from lower layers and result is provided to object detector model
HR-Net (Sun et al., 2019b)	The high-resolution features in their initial architecture for pose estimation (Sun et al., 2019a) is concatenated with features gathered from facial detection and segmentation transformer (Sun et al., 2019b)

Table 3. Baseline Experimental Results

Model	Backbone	Batch	Epoch	Accuracy	IOU	FI
UNet (Ronneberger et al., 2015)	Vgg19 (Simonyan & Zisserman, 2015)	32	50	0.52	0.32	0.39
UNet (Ronneberger et al., 2015)	InceptionV3 (Szegedy et al., 2017)	32	50	0.64	0.37	0.44
UNet (Ronneberger et al., 2015)	ResNet50 (He et al., 2015)	32	50	0.55	0.33	0.42
FPN (Lin et al., 2017)	Vgg19 (Simonyan & Zisserman, 2015)	32	50	0.68	0.72	0.78
FPN (Lin et al., 2017)	InceptionV3 (Szegedy et al., 2017)	32	50	**0.71**	0.81	0.84
FPN (Lin et al., 2017)	ResNet50 (He et al., 2015)	32	50	0.67	0.76	0.82
HR-Net (Sun et al., 2019b)	Vgg19 (Simonyan & Zisserman, 2015)	32	50	0.61	0.76	0.82
HR-Net (Sun et al., 2019b)	InceptionV3 (Szegedy et al., 2017)	32	50	0.68	**0.83**	**0.87**
HR-Net (Sun et al., 2019b)	ResNet50 (He et al., 2015)	32	50	0.65	0.81	0.85

to both false positives and false negatives. Consider a binary classification problem in an image, whose contents are highly imbalanced containing 90% of one class and just 10% of second class. The prediction accuracy will be high for the class that is imbalanced whereas the actual accuracy for all classes can be low.

IOU (Intersection-Over-Union), also referred to as Jaccard Index, is the most commonly used evaluation metrics in domains like Object detection and semantic segmentation. It evaluates the performance of the model by calculating the area of overlap between ground truth results and prediction and area of union. IOU can be used to evaluate multi-classification problems by calculating the IOU of each class and then taking an average.

FI Score is a similar metric to IOU and is generally called by its second name `dice coefficient`. The metric can be calculated using formula `(2 * area overlapped) / (Sum of pixels in the ground truth and predicted)`. Evaluations performed by both IOU and F1 score are correlated. Hence, if one metric gives an indication that the accuracy is low, the other metric will give the same indication. Dice coefficient measurements tend to be closer to average whereas IOU measurements tend towards worst-case performance values.

5. Results and Discussions

We have used transfer learning approach to create a benchmark on proposed dataset. The models used for segmentation training are UNet (Ronneberger et al., 2015), FPN (Lin et al., 2017) and HR-Net (Sun et al., 2019b). A description of models is given in Table 2. To extract feature maps from images and their corresponding masks, multiple classification models' weight sets are used as a backbone, namely, Inception (Szegedy et al., 2017), Vgg19 (Simonyan & Zisserman, 2015), Resnet50 (He et al., 2015) weights set (trained on Imagenet dataset (Deng et al., 2009)).

For training purposes, the training batch size is set to 32 as empirical evidence suggests a batch size of 32 provides the best results (Mishkin et al., 2017). Because of limited computing power, the epoch is set to 50 to get the most results. For the loss function, the Cross-entropy loss is opted for accuracy metric whereas, for IOU and F1 metrics, Jaccard loss and Dice loss are used respectively.

The training results acquired from all models are described in Table 3. Inception architecture as a backbone gives better results as compared with all other backbones for all metrics because of its deep layered architecture. The overall accuracy has behaved differently than other metrics, this is because shrub and vegetation percentage in a single image exceeds any other class percentage in the same image. However, other metrics are used to address this issue which can be evident from the results of HR-Net out-performing other models for both F1 score and IOU metrics.

6. Conclusion and Further Work

This paper introduces a new dataset for forest terrain identification using semantic segmentation, near beaver habitat, from UAV captured images. Nine different classes were identified and annotated using 3rd party tools. Furthermore, the dataset was used to train several state-of-the-art deep

neural networks and a benchmark was provided and analysed for future case studies. For future work, we expect to add more classes and to expand the size of the existing dataset for training larger deep models. Furthermore, we plan to enhance existing deep models to provide a state-of-the-art novel framework that will outperform the existing deep models on our dataset.

References

Achanta, R. and Susstrunk, S. Superpixels and polygons using simple non-iterative clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4651–4660, 2017.

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Süsstrunk, S. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012.

ArcGis. Arcgis online, 2020. URL http://www.arcgis.com.

Arun, P. V., Herrmann, I., Budhiraju, K. M., and Karnieli, A. Convolutional network architectures for super-resolution/sub-pixel mapping of drone-derived images. *Pattern recognition*, 88:431–446, 2019.

Badrinarayanan, V., Kendall, A., and Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation, 2016.

Breiman, L. Random forests. *Machine Learning*, 45 (1):5–32, 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324. URL http://dx.doi.org/10.1023/A%3A1010933404324.

Chen, Y., Ming, D., and Lv, X. Superpixel based land cover classification of vhr satellite image combining multi-scale cnn and scale parameter estimation. *Earth Science Informatics*, 12(3):341–363, 2019.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

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

Everingham, M., Eslami, S. M. A., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, January 2015.

Fikri, M. Y., Azzarkhiyah, K., Al Firdaus, M. J., Winarto, T. A., Syai'in, M., Adhitya, R. Y., Endrasmono, J., Rahmat, M. B., Setiyoko, A. S., Zuliari, E. A., et al. Clustering green openspace using uav (unmanned aerial vehicle) with cnn (convolutional neural network). In *2019 International Symposium on Electronics and Smart Devices (ISESD)*, pp. 1–5. IEEE, 2019.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition, 2015.

Kattenborn, T., Leitloff, J., Schiefer, F., and Hinz, S. Review on convolutional neural networks (cnn) in vegetation remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 173:24–49, 2021.

LabelBox. Labelbox, 2020. URL http://www.labelbox.com.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pp. 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1-55860-778-1. URL http://dl.acm.org/citation.cfm?id=645530.655813.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. Feature pyramid networks for object detection, 2017.

Lyu, Y., Vosselman, G., Xia, G.-S., Yilmaz, A., and Yang, M. Y. Uavid: A semantic segmentation dataset for uav imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 165:108 – 119, 2020. ISSN 0924-2716. doi: https://doi.org/10.1016/j.isprsjprs.2020.05.009. URL http://www.sciencedirect.com/science/article/pii/S0924271620301295.

Mishkin, D., Sergievskiy, N., and Matas, J. Systematic evaluation of convolution neural network advances on the imagenet. *Computer Vision and Image Understanding*, 161:11–19, 2017.

Nummi, P., Vehkaoja, M., Pumpanen, J., and Ojala, A. Beavers affect carbon biogeochemistry: both short-term and long-term processes are involved. *Mammal Review*, 48(4):298–311, 2018. doi: 10.1111/mam.12134.

QGIS. Qgis, 2020. URL https://www.qgis.org/en/site/.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation, 2015.

Shelhamer, E., Long, J., and Darrell, T. Fully convolutional networks for semantic segmentation, 2016.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition, 2015.

Sun, K., Xiao, B., Liu, D., and Wang, J. Deep high-resolution representation learning for human pose estimation, 2019a.

Sun, K., Zhao, Y., Jiang, B., Cheng, T., Xiao, B., Liu, D., Mu, Y., Wang, X., Liu, W., and Wang, J. High-resolution representations for labeling pixels and regions, 2019b.

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, pp. 4278–4284. AAAI Press, 2017.

Wang, S., Liu, L., Qu, L., Yu, C., Sun, Y., Gao, F., and Dong, J. Accurate ulva prolifera regions extraction of uav images with superpixel and cnns for ocean environment monitoring. *Neurocomputing*, 348:158–168, 2019.

Yang, F., Sun, Q., Jin, H., and Zhou, Z. Superpixel segmentation with fully convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13964–13973, 2020.

Zürn, J., Burgard, W., and Valada, A. Self-supervised visual terrain¨ classification from unsupervised acoustic feature learning. *IEEE Transactions on Robotics*, 2020.