Vegetation Detection in UAV Imagery for Railway Monitoring

Md Atiqur Rahman and Abdelhamid Mammeri
National Research Council Canada, Ottawa, Canada

Keywords: Vegetation Detection on Railway Tracks, Drone-based Vegetation Detection, Image Semantic Segmentation.

Abstract: Vegetation management on and alongside the railway tracks is very crucial for safe railway operations. The railway industry, therefore, needs to regularly monitor the growth of vegetation on railway tracks and embankments and mostly relies on human inspectors for the inspection and monitoring. This manual process being prohibitively time-consuming and cost-ineffective, there is a growing need to automate the process of vegetation detection. Aerial imagery collected using Unmanned Aerial Vehicles (UAVs) is becoming increasingly popular for automated inspection and monitoring. On the other hand, due to their recent success, Deep Convolutional Neural Networks (DCNNs) have seen rapid deployment in a wide array of image understanding tasks. In this work, we therefore, investigate the effectiveness of DCNNs for automating the vegetation detection task using UAV imagery. We further propose simple yet effective modification to an existing DCNN architecture and demonstrate its efficacy for vegetation detection using publicly available dataset.

1 INTRODUCTION

Growth of vegetation on and alongside the railway tracks presents potential hazards associated with the railway operations and challenges the ability of engineering forces to maintain safe track conditions. Apart from causing hindrance to proper inspection of the track structure and trains (CN, 2019), growth of vegetation in the trackbed can clog the ballast causing poor track drainage which may eventually lead to the collapse of the railway embankment (Briggs, 2010; Scott et al., 2007). It may also increase the braking distance of the trains as it makes the tracks slippery (Nyberg, 2016b). Therefore, the railway industry needs to regularly monitor, detect and control the growth of vegetation in the ballast section and the Rights-of-Way (ROW) to maintain safe functioning of the train operations. Fig. 1 gives an overview of the different zones around the rail tracks where vegetation management is needed.

In its current state, the railway industry mainly relies on human inspectors who walk along the tracks and judge for themselves the extent and condition of vegetation growth on a regular basis. Apart from being expensive, such manual inspection is very time-consuming specially for the Canadian railway network that stretches from east coast to west coast. Therefore, automating the process of vegetation detection on and alongside the railway tracks is of utmost interest to the railway industry.

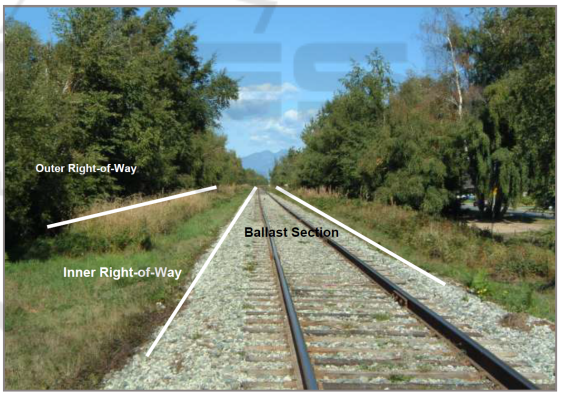

Figure 1: Different zones on and around the rail tracks where vegetation management is needed. Image source CN (2019).

The very limited body of research (e.g., (Nyberg, 2016b; Yella et al., 2013; Nyberg et al., 2013)) that delves into automating (or, semi-automating) the process of vegetation detection along railway tracks is based on acquiring track images using service locomotives equipped with trolley-mounted cameras. However, such ground-based imagery has limited Field of View (FOV) which may not be adequate enough to capture all the Rights-of-Way as depicted in Fig. 1. Besides, data collection for these methods requires occupying the tracks, and therefore, could hamper the regular train operations. Moreover, the existing methods mostly rely on machine vision tech-

Rahman, M. and Mammeri, A.
Vegetation Detection in UAV Imagery for Railway Monitoring.
DOI: 10.5220/0010439904570464
In *Proceedings of the 7th International Conference on Vehicle Technology and Intelligent Transport Systems (VEHITS 2021)*, pages 457-464
ISBN: 978-989-758-513-5

niques and have not explored the effectiveness of the recent sophisticated deep learning based image understanding algorithms, such as the Deep Convolutional Neural Networks (DCNNs).

Compared to locomotive-mounted imagery, images captured using Unmanned Aerial Vehicles (UAVs), or drones as they are more commonly known, have very large FOV allowing one to analyse large areas with sufficient surrounding context. Additionally, drones offer flexible flying schedules without obstructing regular train operations. On the other hand, drone imagery provides substantially higher level of details compared to other modes of aerial imagery (e.g., satellites). The geometric resolution of drone imagery lies in the range of **2–5 cm** as opposed to **50–100 cm** for satellite imagery. Drones therefore provide for a more economical, fast and efficient approach to vegetation detection along the rail tracks.

This paper, therefore, focuses on vegetation detection in drone imagery. We are particularly inspired by the recent success of the DCNNs that have already proved to be a very powerful machine learning technique for a variety of image understanding tasks including pixel-wise image semantic segmentation (Seferbekov et al., 2018; Chen et al., 2018; Shelhamer et al., 2017). To this end, we review some of the recent DCNN architectures proposed for semantic image segmentation and evaluate their efficacy for vegetation detection in drone imagery. Additionally, we propose simple yet effective modification to an existing DCNN architecture that is shown to produce better results on vegetation detection.

2 RELATED WORKS

2.1 Vegetation Detection on Railway Tracks

To the best of our knowledge, there has been no research conducted on vegetation detection on railway tracks using drone-based imagery. The limited body of research that addresses the problem is based on ground-based imagery. An early work (Hulin and Schussler, 2005) collected multi-spectral images along railway tracks using locomotive-mounted cameras and performed detection and measurement of vegetation using spectral analysis. Other works (Yella et al., 2013; Nyberg et al., 2013) applied machine vision techniques, such as color segmentation and morphological operations to segment vegetation pixels on the railway embankments. Roger Nyberg, in his seminal thesis (Nyberg, 2016b), showed that machine vision algorithms are able to produce satisfactory results in quantifying the vegetation cover and classifying the plant species as well. The same author in a later work (Nyberg, 2016a) applied classical machine learning approach (e.g., Bag-of-Features model) to classify woody plants on railway trackbeds and embankments. However, none of these works are based on drone-based imagery nor explored the recent advanced machine learning algorithms such as, DCNNs which is what motivates this work.

2.2 Drones for Visual Recognition

Though initially designed for military use, thanks to their low-cost, flexible operations, and maneuverability drones are now being used for commercial purposes. The abundance of visual data collected using drone imagery coupled with the recent sophisticated deep learning based algorithms has seen rapid deployment in a wide range of applications. This includes precision agriculture for crop monitoring (Duarte-Carvajalino et al., 2018), crop yield estimation (Wahab et al., 2018), and weed mapping (Huang et al., 2018); security surveillance for person re-identification (Grigorev et al., 2020), crowd counting (Küchhold et al., 2018), and vehicle tracking (Song et al., 2020); and search and rescue operations (Bejiga et al., 2017; Quan et al., 2019) for disaster and crisis management.

The railway industry has recently started to capitalize drone imagery coupled with vision-based methods mainly for track extraction (Singh et al., 2019), examining high-voltage electrical lines (Clark, 2020), as well as inspecting railway catenary lines and the alignment of tracks and switching points (THALES, 2019). However, drone-based vegetation detection on railway tracks has not yet been explored. Therefore, in this paper, we aim at detecting vegetation on and alongside the railway tracks using drone imagery.

2.3 Image Semantic Segmentation using DCNN

Given an input image, semantic segmentation aims at assigning a class label to each pixel. The recent advancement in semantic segmentation has been mainly driven by the huge success of the DCNNs. Long *et al.*, in his pioneering work (Shelhamer et al., 2017), proposed a Fully Convolutional Network (FCN) that has now become a de-facto DCNN architecture for the recent semantic segmentation methods. Based on FCN, Ronneberger et al. (2015) proposed a network architecture called U-Net consisting of a contracting path and a symmetric expanding path to capture the

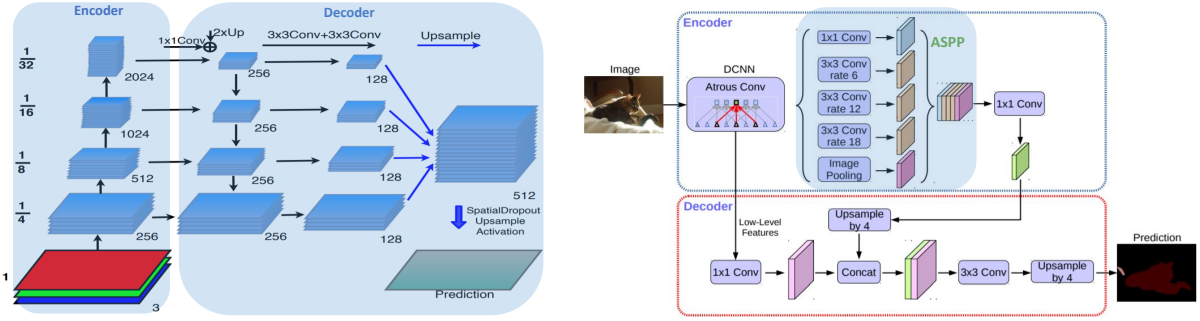

Figure 2: Architecture of FPN for segmentation (Seferbekov et al., 2018) (left) and DeepLabv3+ (Chen et al., 2018) (right). Best viewed when zoomed in.

context, thereby achieving precise localization. To handle multiple scales of objects, Seferbekov et al. (2018) proposed a segmentation model that makes use of the Feature Pyramid Network (Lin et al., 2017) that capitalizes on multi-scale feature hierarchy. Chen et al. (2017), on the other hand, proposed DeepLabv3 employing Atrous Spatial Pyramid Pooling (ASPP) module based on dilated convolutions to capture features at different scales. DeepLabv3+ (Chen et al., 2018) further enhanced this architecture by introducing lateral connections in the decoder module and enjoys the current state-of-the-art results on several segmentation benchmarks.

3 METHODOLOGY

Since vegetation does not have regular shape and structure, the vegetation detection problem is usually formulated as an image semantic segmentation problem, where the goal is to label the vegetation pixels in the input image. In the following sections, we describe two state-of-the-art DCNN architectures for pixel-wise image semantic segmentation followed by our proposed modified DCNN.

3.1 DCNN for Image Semantic Segmentation

For the purpose of this work, we choose two state-of-the-art semantic segmentation networks, namely, Feature Pyramid Network (FPN) for segmentation (Seferbekov et al., 2018) and DeepLabv3+ (Chen et al., 2018). The former won the DEEPGLOBE–CVPR 2018 land cover segmentation challenge from satellite imagery, thus highly relevant to the task at hand. On the other hand, DeepLabv3+ is the current state-of-the-art method on several semantic segmentation benchmarks.

3.1.1 Feature Pyramid Network (FPN) for Segmentation

Fig. 2 (left) shows the architecture of the network proposed in Seferbekov et al. (2018) that is based on the Feature Pyramid Network (FPN) (Lin et al., 2017). The network is composed of an encoder and a decoder. The encoder is basically having a bottom-up pathway that receives an RGB image as input and progressively downsamples the image in spatial dimensions using strided convolutions. The decoder, on the other hand, follows a top-down pathway to progressively upsample the feature maps using bilinear interpolation with lateral connections from the correspondingly sized feature maps in the encoder module. Once the feature pyramid is built, the channel dimension of the feature maps are reduced by using 3×3 convolutions and then upsampled using bilinear interpolation to match the size of the finest resolution feature map at the bottom. Finally, all these maps are concatenated together along the channel dimension followed by 1×1 convolution and spatial dropout to produce class predictions which are again upsampled to produce an output equal to the size of the input.

3.1.2 DeepLabv3+

The architecture of the DeepLabv3+ network is shown in Fig. 2 (right). Unlike FPN, the encoder of DeepLabv3+ is based on ASPP module that uses atrous convolutions at different dilation rates (e.g., 6, 12, 18) along with image-level feature pooling to capture multi-scale features. These multi-scale feature maps are then concatenated together followed by upsampling by a factor of 4 and then concatenated with a higher resolution feature map extracted from the bottom part of the encoder module. Finally, class predictions are made on this combined feature map followed by bilinear upsampling of the predictions to produce the output.

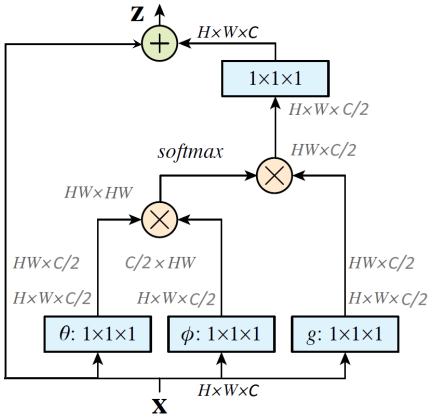

Figure 3: Architecture of non-local neural network (Wang et al., 2018). ⊕ and ⊗ represent element-wise addition and element-wise multiplication, respectively.

3.2 Non-local Neural Network (NNN) for Semantic Segmentation

In this section, we describe the proposed modification that is based on a novel work called Non-local Neural Network (NNN) (Wang et al., 2018). Although the original work was proposed for video classification, we repurpose it for the segmentation task by incorporating NNN in the FPN architecture. As aptly reasoned in Wang et al. (2018), capturing long-range dependencies is very important for any deep neural networks. Since the convolution operation can only capture local neighborhoods, we include NNN blocks to allow the network to better capture the subtle contexts that may be present far from the current spatial position.

An NNN block, as shown in Fig. 3, basically tries to capture long range dependencies by computing the response at a position as a weighted sum of the features at all spatial positions in the input feature maps. Mathematically, the non-local operation can be formulated by the following equation.

$$y_i = \frac{1}{C(x)} \sum_{\forall_j} f(x_i, x_j) g(x_j) \tag{1}$$

where x is the input, y is the output, i is the index of the output spatial position, j is the index that enumerates all possible input spatial positions, and $C(x)$ is a normalization factor. The pairwise function f can be formulated as follows.

$$f(x_i, x_j) = e^{\theta(x_i)^T \phi(x_j)} \tag{2}$$

Here, θ and ϕ are two parameterized embedding functions, such as $\theta(x_i) = W_\theta x_i$ and $\phi(x_j) = W_\phi x_j$.

FPN with Non-local Neural Networks: Fig. 4 shows our proposed modification to the FPN architecture. Instead of concatenating the final feature maps,

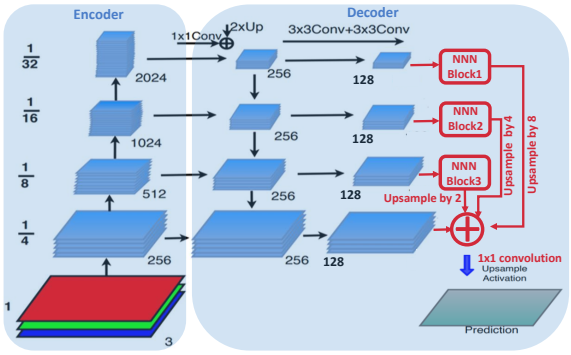

Figure 4: Modified architecture of FPN for segmentation. We use non-local blocks on the top three feature maps and combine all maps using element-wise addition. Best viewed when zoomed in.

we pass the top three maps through individual NNN blocks. For computational efficiency, we do not apply NNN on the finest resolution feature map. The output of each of the NNN blocks are upsampled to match the size of the bottom-most feature map and combined with the bottom-most feature map using element-wise addition. Finally, predictions are made on the combined feature map followed by upsampling (by a factor of 4) to match the size of the input.

4 EXPERIMENTAL SETUP

Currently, there is no benchmark publicly available for vegetation detection on and alongside the railway tracks using drone imagery. To circumvent this, we evaluate the different methods discussed in Section 3 based on a publicly available benchmark for semantic segmentation in drone imagery that depicts urban scenario. We conjecture that methods that are successful for vegetation detection in urban scenario can be readily adapted to the rail context. This is supported by the fact that rail tracks mainly pass through rural and remote areas that have much less scene complexity compared to the urban setting.

4.1 Dataset

We use a publicly available dataset called 'Semantic Drone Dataset' (Graz University of Technology, 2020) that contains drone-imagery depicting urban scenes having different objects including vegetation. Fig. 5 shows an example image and the corresponding annotation from the dataset.

The dataset contains 400 very high resolution (4000×6000 pixels) drone imagery with polygonal coarse annotations for 22 different object categories including 'bald-tree', 'grass', 'tree' and 'other-

vegetation'. In this work, we are only interested in these 4 classes while we consider the other classes as background. The images are captured from nadir view at an altitude of 5–30 m. Since the test set is not publicly available, we split the 400 images into train:test:validation sets following a 300:70:30 random split, respectively.

4.2 Training Configurations

We first resize the images to 1000×1500 pixels for faster inference. To increase the amount of training data, we perform various data augmentations including affine transformations (e.g., scaling, rotation, and shifting), color distortion and addition of Gaussian noise. For all the models, we use the ResNeXt-50(32x4d) (Xie et al., 2017) as the encoder network pretrained on Imagenet dataset (Russakovsky et al., 2015). Training is performed on random crops of size 512×512 with a batch size of 8, whereas, inference is performed on the whole image (i.e., 1000×1500). We train the different models using Adam optimizer and the categorical cross-entropy loss with class-weighting with an initial learning rate of 0.0001 which is reduced by a factor of 10 every 400 epochs. Models are trained for a total of 1000 epochs, while model selection is performed based on performance on the validation set. All the models are trained using the publicly available library called 'Segmentation Models Pytorch' (Yakubovskiy, 2020).

4.3 Evaluation Metrics

To evaluate the performance of the different methods, we use four performance metrics. The first one is called *mean Intersection over Union (mIoU)* which is usually used as the standard performance measure for multi-class semantic segmentation. Additionally, to evaluate the performance of the binary vegetation segmentation task, we use three other measures – *Precision–Recall curve (PR curve), F-measure* (F_β) *curve*, and *Mean Absolute Error (MAE)*.

4.3.1 Mean Intersection over Union (mIoU)

As defined in Eq. 3, *mIoU* is defined as the mean of the *IoU*'s for the different classes. *IoU*, on the other hand, is defined as the intersection over union between the ground-truth and the predicted object regions. Equation 4 defines IoU.

$$mIoU = \frac{1}{C}\sum_{i=1}^{C} IoU_i \tag{3}$$

$$IoU = \frac{TP}{TP+FP+FN} \tag{4}$$

Figure 5: Sample image (left) and the corresponding annotation (right) from the 'Semantic Drone Dataset' (Graz University of Technology, 2020).

Here, C denotes the total number classes, whereas, TP (true-positive), FP (false-positive) and FN (false-negative) denote the total number of pixels correctly predicted as vegetation, incorrectly predicted as vegetation, and incorrectly predicted as background, respectively.

4.3.2 PR Curve

In the context of vegetation segmentation, precision and recall can be defined as the fractions of the predicted and ground-truth vegetation pixels, respectively that are correctly predicted. Equation 5 defines precision and recall. The PR curve can be generated by plotting each (precision, recall) pair while varying the classification decision threshold from 1 to 0.

$$Precision = \frac{TP}{TP+FP} \qquad Recall = \frac{TP}{TP+FN} \tag{5}$$

where, *TP, FP, FN* denote the same as in Eq. 4.

4.3.3 F-measure Curve

As defined in Eq. 6, *F-measure* (F_β) combines the *Precision* and *Recall* into a single value, thus provides a more comprehensive quantitative evaluation. The *F-measure* curve can be generated by plotting the F_β values for each pair of (precision, recall) as the classification decision threshold varies from 1 to 0. Apart from the *F-measure* curve, we also report the maximum of F_β values ($max(F_\beta)$).

$$F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{6}$$

Following Achanta et al. (2009), we set $\beta^2 = 0.3$ to give more weight to *Precision* than *Recall*.

4.3.4 Mean Absolute Error (MAE)

MAE (Perazzi et al., 2012) denotes the average of the per-pixel absolute differences between the ground-truth and the predicted segmentation masks. Given a dataset of N images, where each ground-truth mask

Table 1: Comparison of class-wise *IoU (%)* and *mIoU (%)* (higher is better) on the test split. 'Bkg' stands for background.

	bald-tree	grass	tree	other-vegetation	Bkg	*mIoU*
FPN	86.9	86.5	81.2	65.4	83.1	80.6
DeepLabv3+	87.9	87.6	81.3	66.6	82.9	81.3
FPN w/ NNN	86.6	87.8	82.8	71.4	83.7	**82.5**

Table 2: Confusion matrix for FPN w/ NNN on the test split. An entry at (x,y) denotes the percentage of pixels belonging to object x that is classified as object y. 'Bkg' stands for background.

	bald-tree	grass	tree	other-vegetation	Bkg
bald-tree	**88.4**	3.1	0.0	3.0	5.5
grass	0.6	**95.0**	0.4	1.6	2.4
tree	0.3	0.8	**88.4**	9.5	1.0
other-vegetation	0.7	4.3	6.3	**83.8**	4.9
Bkg	0.2	0.6	0.4	3.2	**95.6**

(G) and the predicted mask (P) have spatial dimensions of $H \times W$, the *MAE* can be defined as follows.

$$MAE = \frac{1}{N \times H \times W} \sum_{n=1}^{N} \sum_{i=1}^{H} \sum_{j=1}^{W} |G^n_{i,j} - P^n_{i,j}| \quad (7)$$

5 RESULTS

5.1 Multi-class Vegetation Segmentation

Table 1 reports class-wise *IoU* and the *mIoU* of FPN, DeepLabv3+ and the proposed variant of FPN (i.e., FPN w/ NNN) across the different vegetation classes. DeepLabv3+ performs better than FPN, which is in agreement with the results on other segmentation benchmarks reported in the literature. As we can see, the proposed variant of FPN performs better than the other two models (1.9 and 1.2 percentage points higher *mIoU* than FPN and DeepLabv3+, respectively), thereby demonstrating its superiority for the multi-class vegetation detection task.

One observation is that, the *IoU* of the class 'other-vegetation' is consistently low across the different models when compared to the other classes. To investigate into this, we show the confusion matrix for the different vegetation classes in Table 2. As the table reveals, the class 'other-vegetation' is confused for 'tree' for 6.3% of the times, whereas, 'tree' is confused for 'other-vegetation' for 9.5% of the times. This clearly indicates that there exists high inter-class similarity between these two classes in the dataset.

Figure 6: Example image (left) and the corresponding annotation (right) from the 'Semantic Drone Dataset' showing high inter-class similarity between the classes 'tree' and 'other-vegetation'. Best viewed when zoomed in.

Table 3: Comparisons of *IoU(%)* (higher is better), $max(F_\beta)$ (higher is better), and *MAE* (lower is better) of the different models on the test split. 'Bkg' stands for background.

	Vegetation	Bkg	*mIoU*	$max(F_\beta)$	*MAE*
FPN	89.7	95.1	92.4	0.930	0.052
DeepLabv3+	91.1	96.4	93.8	0.936	0.047
FPN w/ NNN	92.2	95.6	**93.9**	**0.942**	**0.044**

We can confirm this by visually looking into the sample images and the corresponding annotations for the two classes. As shown in Fig. 6, it is very difficult even for a human to distinguish between these two classes from a nadir view especially with the presence of other artifacts such as, shadow and occlusion from nearby objects.

5.2 Binary Vegetation Segmentation

For the purpose of vegetation detection on and alongside the railway tracks, we do not need to distinguish among the different vegetation classes. Hence, we combine the different vegetation classes together as a single semantic class called 'vegetation' and consider everything else as background. Under this binary setting, Table 3 compares the different models in terms of *IoU*, $max(F_\beta)$, and *MAE*. As the table reveals, the proposed modification is capable of providing performance improvement across the different performance metrics. This clearly demonstrates the efficiency of the proposed modification for the vegetation detection task. As observed in earlier experiments in Section 5.1, DeepLabv3+ performs better than FPN.

However, compared to the multi-class segmentation results as reported in Table 1, the *IoU* values for all the models are improved under the binary segmentation setting. This can be mainly attributed to the fact that, with the merging of the different vegetation classes, the high inter-class similarity that exists between the different vegetation classes (e.g., 'tree' and 'other-vegetation') is diminished as they are now considered the same class.

To further evaluate the quality of the predicted segmentation masks, we plot *PR* and *F-measure*

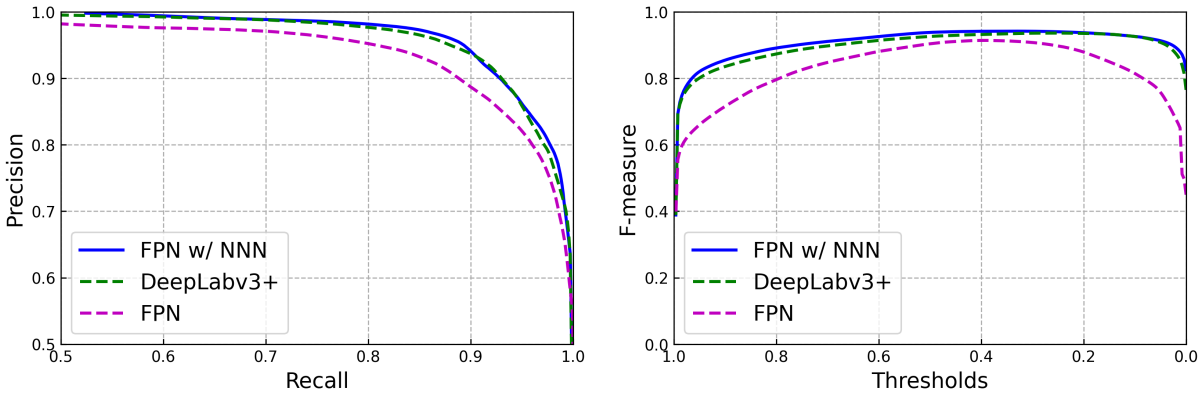

Figure 7: *PR* curves (left) and *F-measure* curves (right) for the different models. The proposed model FPN w/ NNN is capable of producing higher area under the curves than the other models.

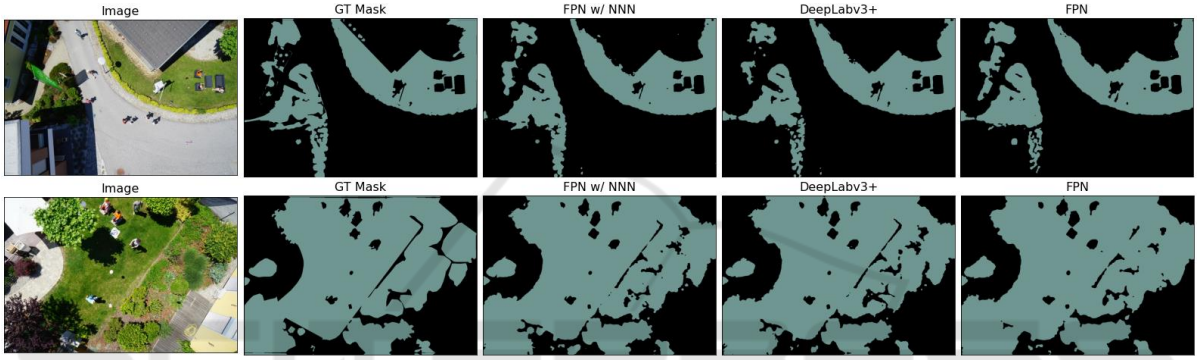

Figure 8: Some qualitative results on vegetation detection using the different models. Each row (from left to right) shows an image from the test split, the ground-truth segmentation mask, and the predicted segmentation masks as produced by FPN w/NNN, DeeLabv3+, and FPN, respectively. Best viewed when zoomed in.

curves for the different models as shown in Fig. 7. The higher area under the curves as achieved by the model FPN w/ NNN clearly indicates its superiority for the binary vegetation segmentation task.

5.3 Qualitative Results

Fig. 8 shows some qualitative results achieved by using the different models. As we can see, the predicted segmentation mask for FPN w/ NNN is qualitatively very close to the ground-truth segmentation mask, thereby indicating its efficacy for the vegetation detection task.

6 CONCLUSION

In this work, we demonstrate the effectiveness of some state-of-the-art deep semantic segmentation models for the task of vegetation detection from UAV imagery. We additionally propose a modified DCNN architecture to further improve the performance of the vegetation detection task. Though the methods are shown to work in an urban setting, they can be readily adapted to vegetation detection on and alongside the railway tracks. As a future work, we aim to collect drone images on the railway tracks and test the methods in a railway context.

REFERENCES

Achanta, R., Hemami, S., Estrada, F., and Susstrunk, S. (2009). Frequency-tuned salient region detection. In *CVPR*.

Bejiga, M. B., Zeggada, A., Nouffidj, A., and Melgani, F. (2017). A convolutional neural network approach for assisting avalanche search and rescue operations with UAV imagery. *Remote Sensing*, 9(2).

Briggs, K. (2010). *Charing Embankment: Climate Change Impacts on Embankment Hydrology*, pages 28–31.

Chen, L., Papandreou, G., Schroff, F., and Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*.

Clark, M. (03 July 2020 (accessed August 17, 2020)). *The Rise Of The Railway Drone*.

CN (2019). Pest management plan for integrated vegetation control, Canadian National Railway Company (CN) pest management plan. 5-yearly plan, BC, Canada.

Duarte-Carvajalino, J., Alzate, D., Ramirez, A., Santa, J., Fajardo-Rojas, A., and Soto-Suárez, M. (2018). Evaluating late blight severity in potato crops using unmanned aerial vehicles and machine learning algorithms. *Remote Sensing*, 10.

Graz University of Technology (2020). Semantic drone dataset. http://dronedataset.icg.tugraz.at. Retrieved 2020-06-01.

Grigorev, A., Liu, S., Tian, Z., Xiong, J., Rho, S., and Feng, J. (2020). Delving deeper in drone-based person re-id by employing deep decision forest and attributes fusion. *ACM Trans. Multimedia Comput. Commun. Appl.*, 16(1).

Huang, H., Deng, J., Lan, Y., Yang, A., Deng, X., and Zhang, L. (2018). A fully convolutional network for weed mapping of unmanned aerial vehicle (UAV) imagery. *PLOS ONE*.

Hulin, B. and Schussler, S. (2005). Measuring vegetation along railway tracks. In *Proc. 2005 IEEE Intelligent Transportation Systems*, pages 561–565.

Küchhold, M., Simon, M., Eiselein, V., and Sikora, T. (2018). Scale-adaptive real-time crowd detection and counting for drone images. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 943–947.

Lin, T., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017). Feature pyramid networks for object detection. In *CVPR*.

Nyberg, R. (2016a). A machine learning approach for recognising woody plants on railway trackbeds. In *Proc. of International Conference on Railway Engineering (ICRE)*.

Nyberg, R., Gupta, N., Yella, S., and Dougherty, M. (2013). Detecting plants on railway embankments. *Journal of Software Engineering and Applications*, pages 8–12.

Nyberg, R. G. (2016b). *Automating Condition Monitoring of Vegetation on Railway Trackbeds and Embankments*. PhD thesis.

Perazzi, F., Krähenbühl, P., Pritch, Y., and Hornung, A. (2012). Saliency filters: Contrast based filtering for salient region detection. In *CVPR*.

Quan, A., Herrmann, C., and Soliman, H. (2019). Project vulture: A prototype for using drones in search and rescue operations. In *2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, pages 619–624.

Ronneberger, O., P.Fischer, and Brox, T. (2015). U-Net: convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Scott, J., Loveridge, F., and O'Brien, A. (2007). Influence of climate and vegetation on railway embankments. In *Geotechnical Engineering in Urban Environments: Proceedings of the 14th European Conference on Soil Mechanics and Geotechnical Engineering*, pages 659–664.

Seferbekov, S., Iglovikov, V., Buslaev, A., and Shvets, A. (2018). Feature pyramid network for multi-class land segmentation. In *CVPR*.

Shelhamer, E., Long, J., and Darrell, T. (2017). Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):640–651.

Singh, A. K., Swarup, A., Agarwal, A., and Singh, D. (2019). Vision based rail track extraction and monitoring through drone imagery. *ICT Express*, 5(4):250 – 255.

Song, W., Li, S., Chang, T., Hao, A., Zhao, Q., and Qin, H. (2020). Cross-view contextual relation transferred network for unsupervised vehicle tracking in drone videos. In *WACV*.

THALES (2019). How drones will change the future of railways. https://www.thalesgroup.com/en/worldwide/transport/magazine/how-drones-will-change-future-railways. Retrieved 2020-08-17.

Wahab, I., Hall, O., and Jirström, M. (2018). Remote sensing of yields: Application of UAV imagery-derived NDVI for estimating maize vigor and yields in complex farming systems in sub-saharan Africa. *Drones*, 2(3):28.

Wang, X., Girshick, R., Gupta, A., and He, K. (2018). Non-local neural networks. In *CVPR*.

Xie, S., Girshick, R., Dollar, P., Tu, Z., and He, K. (2017). Aggregated residual transformations for deep neural networks. In *CVPR*.

Yakubovskiy, P. (2020). Segmentation models pytorch. https://github.com/qubvel/segmentation_models.pytorch.

Yella, S., Nyberg, R., Payvar, B., Dougherty, M., and Gupta, N. (2013). Machine vision approach for automating vegetation detection on railway tracks. *Journal of Intelligent Systems*, pages 179–196.