This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2022.3203145

Deep Semantic Segmentation of Trees Using Multispectral Images

Irem Ulku, Erdem Akagündüz *Member, IEEE,* and Pedram Ghamisi *Senior Member, IEEE,*

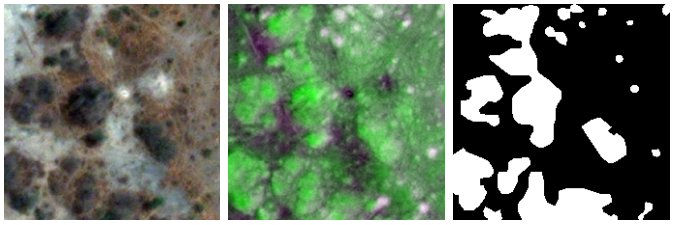
***Abstract*—Forests can be efficiently monitored by automatic semantic segmentation of trees using satellite and/or aerial images. Still, several challenges can make the problem difficult, including the varying spectral signature of different trees, lack of sufficient labelled data, and geometrical occlusions. In this paper, we address the tree segmentation problem using multispectral imagery. While we carry out large-scale experiments on several deep learning architectures using various spectral input combinations, we also attempt to explore whether hand-crafted spectral vegetation indices can improve the performance of deep learning models in the segmentation of trees. Our experiments include benchmarking a variety of multispectral remote sensing image sets, deep semantic segmentation architectures, and various spectral bands as inputs, including a number of hand-crafted spectral vegetation indices. From our large-scale experiments, we draw several useful conclusions. One particularly important conclusion is that, with no additional computation burden, combining different categories of multispectral vegetation indices, such as NVDI, ARVI, and SAVI, within a single three-channel input, and using the state-of-the-art semantic segmentation architectures, tree segmentation accuracy can be improved under certain conditions, compared to using high-resolution visible and/or near-infrared input.**

***Index Terms*—Satellite imagery, semantic segmentation, tree segmentation, vegetation indices**

I. Introduction

FORESTS are one of the principal factors that maintain Earth's climate stability [1]. Accurate monitoring of their state and sustainability is a global concern. Compared to other more traditional methods such as aerial surveys or plot-based analyses, remote sensing has proven to be the most efficient way to monitor forest cover change processes. Conventionally, remote monitoring of processes like deforestation or forest degradation has been interpreted manually by expert analysts. However, with recent developments in the field of artificial intelligence (AI), deep learning (DL) algorithms using satellite and/or aerial imagery are becoming the de facto tool in forest monitoring [2].

Remote sensing imagery can be categorized by the altitude of the aircraft/spacecraft or the type of the optical system (or, simply put, the sensor) [3]. These two factors largely determine the trade-off between the area covered and the amount of detail (i.e., the resolution) of the constructed image. Depending on the type of the sensor, the optical system can sense a part of the electromagnetic spectrum, sample different spectral bands separately and simultaneously, and hence create multi- or hyperspectral imagery. In addition, various spectral bands are used to construct the so-called "spectral indices," which are mathematical combinations of spectral reflectance of different wavelengths. These hand-crafted indices are used to detect the presence of objects or situations, such as vegetation [4], water [5], fire [6], and landslide [7]. Although these hand-crafted indices are valuable features for pattern recognition, today's trend is moving toward building non-transparent, end-to-end deep learning models to detect any feature of interest, usually using supervised techniques [8].



(a) RGB (b) Vegetation (c) Ground Truth

Fig. 1 Example of tree cover in satellite imagery. a) RGB image. b) NVDI, ARVI, and SAVI vegetation indices depicted in three-channel false colour. c) Tree cover ground truth for semantic segmentation.

In this paper, we address the problem of segmenting forest areas (or simply trees) in both satellite and aerial images using deep semantic segmentation techniques and multispectral imagery. We attack the problem in multiple dimensions. To begin with, we examine the effect of altitude using satellite and aerial images with various metric and spectral resolutions. Secondly, we benchmark different deep semantic segmentation architectures, including models with pre-trained convolutional encoders. We utilize different spectral bands, such as visible (RGB), near-infrared (NIR), and short-wave infrared (SWIR) (and their multispectral combinations), as raw inputs to these models. Moreover, we experiment with hand-crafted spectral vegetation indices as input and analyze their performance compared to raw spectral input images. Our goal is to determine how to efficiently utilize deep learning-based architectures for tree segmentation, what practical issues arise with limited remote sensing data, and which methods should be utilized. In order to accomplish this, we conduct large-scale experiments.

The remainder of the paper is organized as follows: in the

Irem Ulku is with the Department of Computer Engineering, Ankara University, Turkey (e-mail: irem.ulku@ankara.edu.tr).

Erdem Akagündüz is with the Graduate School of Informatics, Middle East Technical University (METU), Turkey (e-mail: akaerdem@metu.edu.tr).

Pedram Ghamisi is with (1) the Institute of Advanced Research in Artificial Intelligence (IARAI), 1030 Vienna, Austria, and (2) Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Helmholtz Institute Freiberg for Resource Technology, 09599 Freiberg, Germany (e-mail: p.ghamisi@gmail.com).

Manuscript received ??? ??, 2021; revised ??? ??, 2022.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2022.3203145

following section, we refer to the literature on the subject and underline our contribution. The third section provides the details of our benchmarking environment with respect to the image sets, the segmentation models, and the spectral indices we utilize in our experiment. Section 4 presents the experimental results and covers related analyses and discussions. Finally, Section 5 discusses our conclusions.

II. Related Literature

Survey studies on object detection in optical remote sensing images [9]–[11] show that efforts to detect trees in satellite images date back to the early 90s [12]–[14]. Previously, traditional machine learning methods such as multinomial logistic regression (MLR) [15], random forest (RF) [16]–[18], and support vector machine (SVM) [19]–[21] dominated the area of tree detection and segmentation within remote imaging. These studies relied not only on hand-crafted feature extractions but also on feature encoding [22] and feature pooling [23]. Because of the low-level characteristics of the extracted features, these methods were not able to capture high-level features that reside in complex high-dimensional (spatial, temporal, and spectral) satellite or aerial imagery. Recently, deep learning approaches have emerged to tackle such limitations by making use of hierarchical learning processes while extracting high-level, complex abstractions from the data [24]–[29]. Convolutional neural networks (CNNs) [30], with their proven ability to learn these high-level abstract feature representations with the help of convolutional and pooling layers, are the most commonly used deep learning models for semantic segmentation.

Following the recent interest in deep semantic segmentation architectures [31], image context exploitation and fusing features from different stages for image segmentation have given rise to various methods. Pioneered by the fully convolutional network (FCN) [32] and followed by many improvements [33]–[38], the encoder-decoder (ED) architectures show promising success in solving semantic segmentation tasks. In fact, the ED-based CNN architectures are utilized for semantic segmentation of Earth observation data [37], [39]–[45], allowing the fusion of high- and low-level feature maps while improving spatial accuracy in semantic segmentation.

It is also worth mentioning that classical tree segmentation methods commonly utilized 3D airborne light detection and/or range (LIDAR) data [46]–[55]. For instance, when using 3D LIDAR data to perform tree segmentation, data-specific methods that relied on crafted features were employed such as canopy height model [56], based watershed algorithm [57] or normalized point cloud-based layer stacking algorithm [58]. Besides, in some studies both LIDAR and multispectral data were utilized by applying different algorithms for each data type. While implementing machine/deep learning algorithms that are well-suited to multispectral data, the studies concurrently applied methods native to LIDAR data [59]–[61]. Nevertheless, due to its flexibility in both spatial and spectral resolution, and its compatibility with different remote sensing systems (UAV, satellite, etc.), multispectral and/or hyperspectral imagery are today's main standards for not only tree segmentation, but for several other related problems such as crop, water, and field segmentation [3].

The specific problem of detecting trees presents some challenges, such as geometrical complexity, colour variations, and/or self-occlusion of the tree branches [62], just to name a few. As a result of these challenges, the problem proves to be considerably difficult. For example, regarding self-occlusion, even a slight change in the viewpoint could cause the same structures to appear significantly different. The tree segmentation problem is also challenging due to the environmental background, which may contain other types of green vegetation such as small bushes. A closely related problem is differentiating between the view of a large canopy, hence the overlap between adjacent individual trees. A satellite-retrieved spectral signature is the combination of different reflectance sources, such as tree crown, tree crown shadow, soil, and herbaceous vegetation [17]. Another key factor that limits individual tree mapping is that the spatial resolutions of most satellite/aerial sensors are usually not capable of capturing objects with sizes less than one square meter. However, thanks to commercial satellites and UAVs with multispectral cameras, the improved spatial resolution makes it possible to map every tree on large scale [63].

Instead of using individual spectral bands, vegetation indices (VIs) have been proposed in the literature [64]. VIs are derived from varying mathematical combinations of two or more spectral reflectance values to examine the presence of the vegetation within a pixel. VIs have several drawbacks, notably their inadequacy at interpreting the information content for forest canopy, thereby insufficiently quantifying the effects among different variables, and hence making them less favourable for rapidly changing biophysical factors such as chlorophyll content per unit leaf area, leaf angles, and fractional cover [65]. Performances of VIs are compared to that of the spectral mixture analysis (SMA) in terms of tree canopy cover and are found to be poor [66]. Since there is a high correlation between many VIs, they can be used together and complement each other in different aspects by addressing specific needs that characterize tree canopy.

Most of the DL architectures that process satellite/aerial imagery for the segmentation of objects, including trees, are end-to-end models that take multi- or hyperspectral bands as input channels [67]. However, due to both their multimodal, geolocated, and multitemporal nature and the insufficient amount of labelled remote-sensing data, applying deep learning models tends to be a challenging task [25]. Improvements in the high-performance computing architectures such as GPUs lag far behind the growth of high-dimensional big satellite/aerial data; ergo, recent efforts toward building large-scale deep learning models fall short of training big remote-sensing data, unless a breakthrough is achieved in the acceleration of computing power. One solution to this problem, which we propose in this paper, is to utilize VIs as a priori knowledge to the deep learning models. VIs by definition, provide insight into spectral characteristics of a surface by taking into account the interaction between electromagnetic radiation and vegetation, and hence, can be considered physics-based modelled features.

Our study explores the effect of fusing vegetation indices as

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2022.3203145

TABLE I
Features of Benchmark Multispectral Semantic Segmentation Image Sets

Features	Datasets	
	DSTL (satellite)	**RIT18** (aerial)
GSD	0.31 m (RGB) 1.24 m (VIS+NIR) 7.5 m (SWIR)	0.047 m (VNIR)
Ground FOV	25 km²	2,88 km²
Sensor	World-View 3	TetraCam MicroMCA6
Year	2017	2017
No. of Spectral Bands	16	6
Visible Subbands	400-452 nm 448-510 nm 518-586 nm 590-630 nm 632-692 nm	485-495 nm 545-555 nm 675-685 nm
NIR Subbands	706-746 nm 772-890 nm 866-954 nm	715-725 nm 795-805 nm 890-910 nm
SWIR Subbands	1195-1225 nm 1550-1590 nm 1640-1680 nm 1710-1750 nm 2145-2185 nm 2185-2225 nm 2235-2285 nm 2295-2365 nm	none
Bit Resolution	11-bits/pixel (Visible & NIR) 14-bits/pixel (SWIR)	10-bits/pixel
No. of 224 x 224 patches	5985	1778

a physical model and various ED architectures as deep learning models in order to perform tree semantic segmentation tasks effectively and accurately. For this purpose, we implement various ED-based semantic segmentation architectures such as U-Net [34], SegNet [35], DeepLabv3+ [36], DLinkNet [37] , and DFANet [38] , as well as a pre-DL era ensemble learning algorithm, namely Random Forest [68]. We utilize specific vegetation indices so as to integrate physics-based models as additional input information to the benchmarked models. A framework is constructed for satellite and aerial images that not only provides a comparison between different ED-based semantic segmentation architectures but also leads to insights about fusing various vegetation indices or spectral bands for the purpose of pixel-wise tree segmentation.

III. Methodology

In this section, we provide a methodological background for the experiments conducted within our study. We commence by introducing the multispectral image sets utilized in our experiments. We then provide details on the benchmarked semantic segmentation architectures of our experiments and the types of input we provide to these models.

A. Multispectral Image Sets

The previous studies in the literature have proven that high spectral resolution is a major factor in solving the tree segmentation problem, as is the requirement of high spatial resolution [16]. The earlier airborne cameras and satellite-borne multispectral sensors provided high spatial resolution, while their spectral bands were limited to Blue (0.45–0.51 μm), Green (0.51–0.58 μm), Red (0.63–0.69 μm), and Near Infrared 1 (0.77–0.90 μm). Because several satellite-borne sensors such as WorldView-2 and WorldView-3 were launched after 2009, numerous other spectral bands that are strongly related to vegetation properties have also been included. One is the Coastal band (0.40–0.45 μm), where the associated reflectance is highly bound up with the chlorophyll content of the vegetation. Another band, the Red Edge (0.71–0.75 μm), leads to intuitive discrimination of healthy trees, which is the primary task in precision agriculture. Aside from the detection of healthy trees, infected trees can also be exposed if the Yellow band (0.59–0.63 μm) is utilized to extract the yellowness of their crowns. Apart from the Near-Infrared 1 band, the Near-Infrared 2 band (0.86–1.04 μm) can provide additional information on the vegetation analysis since it is less influenced by atmospheric conditions [69]. The shortwave infrared (SWIR) band (1.19–2.36 μm) is useful to characterize vegetation water content since it is sensitive to water absorption.

In order to conduct supervised tree segmentation experiments, multispectral image sets with pixel-label ground truth are required. Table I shows detailed information about the two image sets that we use in our conducted experiments. In Table I, ground sampling distance (GSD) refers to the distance between the centres of two adjacent pixels corresponding to their representation in the real world [70]. The improvement of segmentation performance is significantly related to the spatial resolution, with GSD being the most important descriptor [71]. Additionally, ground field-of-view (FOV) refers to the metric area covered by the sensor at the chosen operational altitude of the sensor during image collection. The reader may refer to [72]–[75] for a list of available multispectral image sets with pixel labels. The two image sets used in our experiments are discussed below.

DSTL Satellite Imagery Feature Detection Image Set: The DSTL Satellite Imagery Feature Detection image set is provided by the Defense Science and Technology Laboratory [76], and includes 25 very high-resolution training images. Each image covers an area of 1 km × 1 km of the Earth's surface. In this image set, both 3-band and 16-band formats are provided with different spatial resolutions. The images are captured by WorldView-3, which is a commercial satellite. The WorldView-3 optical system provides very high-resolution satellite imagery with 31 cm panchromatic resolution, 1.24 m multispectral resolution, and 7.5 m short-wave infrared resolution [77]. WorldView-3 satellite captures the same area with different types of satellite images such as a panchromatic band (high-resolution), an RGB band (high-resolution), and a multispectral band (M-band) (lower resolution), and a shortwave infrared band (A-band) (lowest resolution). The advantageous feature of having 11- and 14-bit as image colour depths enables more information from each pixel to be used in a deep learning model. There are a total of 10 labelled object types, including trees.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2022.3203145

RIT-18 (The Hamlin State Beach Park) Aerial Image Set: The RIT-18 aerial image set [73] was acquired in 2017 by mounting the Tetracam MicroMCA6 multispectral sensor on board the DJI-S1000 octocopter. This unmanned aircraft system (UAS) is much cheaper than those based on manned aircraft and satellite systems. RIT-18 has been collected with a ground sample distance of 0.047 m and includes six very high-resolution multispectral bands. Three of them are visible RGB bands, while the other three are NIR bands, with all having very high spatial resolution. Due to their high spatial resolution,[1] the NIR bands introduced by the RIT-18 are strengthening the discriminative power of deep learning architectures, especially for cases where vegetation exists [73].

B. Semantic Segmentation Architectures

In the past decade, studies on different semantic segmentation architectures have shown tremendous development [31]. Mostly driven by industry, the majority of these studies were designed for medical applications or unmanned systems such as driverless cars. As mentioned in Section II, we benchmark different DL-based semantic segmentation architecture and a pre-DL ensemble learning method called the Random Forest algorithm. In this part, we provide details on the semantic segmentation techniques that we utilize in our experiments.

The pioneering deep learning-based semantic segmentation architecture is the FCN [32]. This network is considered to be the ancestor of today's widely used ED architectures in semantic segmentation [31]. Although there are many different architectures and approaches in deep learning-based semantic segmentation literature, it's safe to say that the ED architecture is an off-the-shelf solution to many semantic segmentation problems.

Due to its role in recovering spatial resolution in extracted feature maps, the decoder is especially important to implementing ED architectures for multispectral imagery so that the high-resolution details can be recovered effectively. Given this effect, we considered various ED architectures, while paying attention to the different aspects of decoders. Upsampling layers are one of the most crucial parts of decoders, as they increase the spatial resolution [78]. Therefore, upsampling layers are the basis for selecting the proper architectures in this paper.

Below we explain the five fundamental deep learning-based semantic segmentation architectures that we utilize in our experiments: U-Net [34], SegNet [35], DeepLabv3+ [36], DLinkNet [37], and DFANet [38]. The first two are considered to be the basic ED examples, while the latter three are advanced state-of-the-art models that have been proven to be more successful in today's semantic segmentation challenges [79]. Finally, we provide information about the Random Forest algorithm, which is one of the strongest pixel-wise tree segmentation methods of the pre-DL era.

U-Net [34]: U-Net is the one of the pioneering and most-used ED approaches in the semantic segmentation literature (Fig. 2a). It is notable for its unique architecture, in which the whole feature map is transported from the encoder to the corresponding building block in the decoder via skip connections. Compared to state-of-the-art architectures, the decoder part of U-Net is computationally inefficient since it reconstructs the segmentation map, the size of which is equal to that of the original image through many up-scaling operations between high-level semantic information and precise local information. However, U-Net is still considered to be a de facto solution to any semantic segmentation problem.

The implemented U-Net model is a four-level-depth symmetric ED model. At each depth, two repeated blocks of convolutional layers are followed by a batch normalization operation during training and a rectified linear unit (ReLU) activation function in the experiment itself. Attached to these two blocks, each level is concluded by a 2x2 max-pooling layer (stride 2) and a dropout layer with rate=0.5, hence halving the spatial dimension at each depth. For each depth, the feature channels vary among 64, 128, 256, and 512, which is also symmetrical in the decoder. In the decoder part of the process, instead of the max-pooling layer, there are up-convolution layers that help double the spatial dimensions. For all convolutional layers, kernel, stride, and padding sizes are set as 3x3, 1x1 and 1x1, respectively, thus keeping the spatial dimensions fixed within a block. In order to implement the binary classification of trees, the final decoder block is followed by a single depth 1x1 convolution layer and pixel-wise sigmoid layers, which provides an input-sized single channel map of the detection measure. Skip connections between each depth carry feature activation values, a trademark of the original U-Net architecture.

Among the many different ED architectures, we consider U-Net a fitting choice for benchmarking, given that it is a basic example of using transposed convolution as the up-sampling layer [78].

SegNet [35]: SegNet is very similar to U-Net architecture. The only difference is that, rather than passing the entire feature map of the encoder to the decoder, SegNet uses a very simple yet computationally efficient transfer of information. Instead of unpooled features, maximum pooling indices are passed (Fig. 2a).

In the implemented SegNet model, the architecture and the related parameters are the same as in the implemented U-Net model described above. The only difference is, as the trademark of the SegNet architecture, skip connections carry only the pooling indices, which is why SegNet is considered to be an efficient implementation of the U-Net.

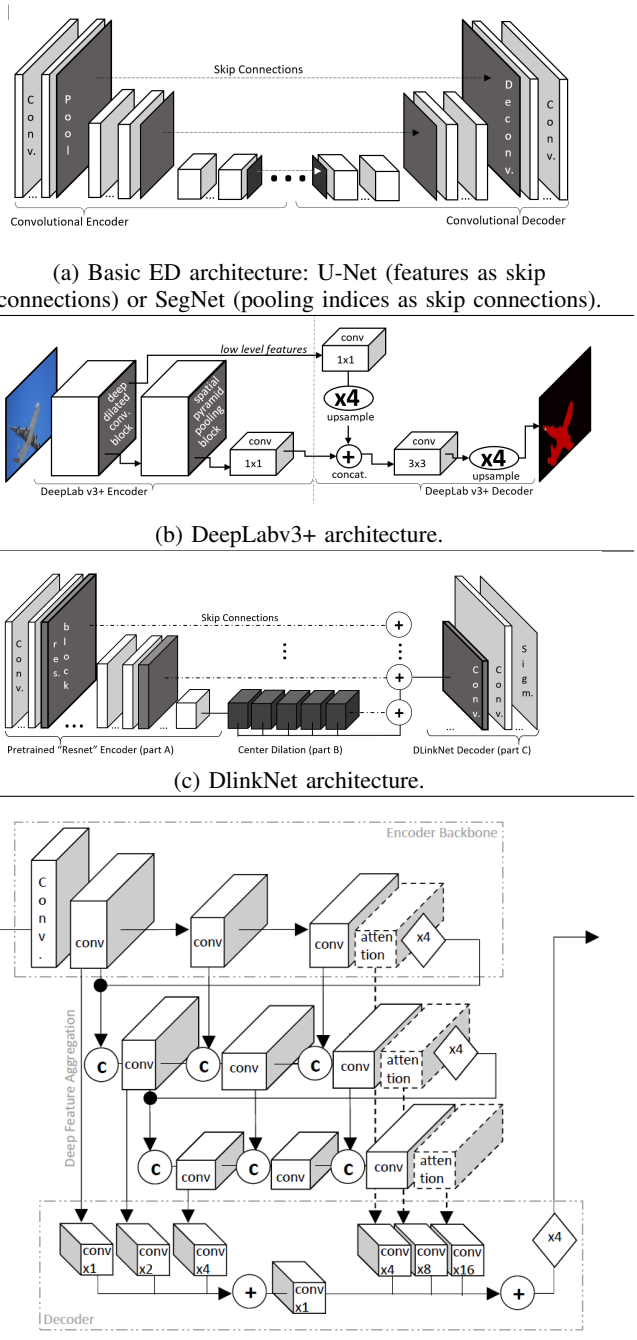
Rather than using the transposed convolution as the upsampling layer like U-Net, SegNet employs interpolation by implementing the unpooling operation [80], [81]. Therefore, the criterion for selecting SegNet is to compare the performance of this upsampling operation at the decoder.

DeepLabv3+ [36]: One of the most advanced, successful, and efficient semantic segmentation frameworks in the literature is the DeepLabv3+ [82] (Fig. 2b), which is the final architecture of the DeepLab family [82]–[84]. It is originally trained with general-purpose image sets such as PASCAL VOC 2012 [79] and COCO [85]. The main idea of DeepLabv3+ is to

[1] 10 and 20 nm band-pass filters were used for this sensor since wider bandwidths tend to create over-saturation [73].

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2022.3203145



(a) Basic ED architecture: U-Net (features as skip connections) or SegNet (pooling indices as skip connections).

(b) DeepLabv3+ architecture.

(c) DlinkNet architecture.

(d) DFANet architecture.

Fig. 2 Different semantic segmentation architectures were utilized in the experiments.

consolidate multi-scale contextual information by using atrous spatial pyramid pooling (ASPP) layers. Compared to a default ED architecture, the decoder part is extremely lightweight, although with better segmentation accuracy.

As semantic segmentation architectures developed rapidly, recent studies have focused on utilizing state-of-the-art architectures such as ResNet [86], Inception [87], and Xception [88] in the encoder, referred to as encoder backbone networks. The goals of this usage are to improve the generalization ability and increase the prediction accuracy of the model. Like many other semantic segmentation architectures, DeepLabv3+ can also function with pre-trained encoders. In scenarios, where the labelled data is limited, different encoder backbone networks may help the model perform with increased efficacy. In the implemented DeepLabv3+ model, the first two levels of the ResNet34 or ResNet101 encoder are transferred for this purpose. However, because we utilized pre-trained encoders that originally accepted three-channel (RGB) input, the experiments with this architecture are limited to spectral input with 3 or fewer channels.

We utilize DeepLabv3+ in our experiments because its upsampling layer is completely different from that of the other ED architectures. DeepLabv3+ employs bilinear upsampling + convolution in its decoder design [78], mainly for computational efficiency.

D-LinkNet [37]: Another state-of-the-art model in the literature is DLinkNet. This model is originally trained with RGB satellite images for pixel-wise road segmentation, which makes it special in the sense that it is one of the rare models specifically designed and trained using image sets with almost-zero zenith angles. The model is based on an approach focused on additional dilated convolution layers in the centre part of an ED architecture (Fig. 2c). DLinkNet is an improvement of the Linknet architecture [89], which is also an ED-based semantic segmentation model.

Similarly to the implemented DeepLabv3+ model, DLinkNet's utilized encoder is transferred and fine-tuned. The first two levels of the ResNet34 encoder is transferred to the implemented DLinkNet model in the conducted experiments. Hence, the experiments with this architecture are also limited to spectral input with less than or equal to 3 channels (please refer to Section III.C for model input types).

DLinkNet is also another example of the ED architecture that relies on transposed convolution in the decoder design [37]. However, it is not considered to be as computationally efficient as the DeepLab family.

DFANet [38]: The majority of recent semantic segmentation studies are targeted at real-time applications like autonomous driving that require fast understanding of the surrounding scenes while achieving high performance [90]–[93]. In our study, we chose the state-of-the-art DFANet as a benchmark model because it has a sophisticated encoder-decoder structure similar to DLinkNet, but is still real-time in performance.

The main contribution of DFANet is the so-called "Cross-level feature aggregation based ED architecture" (Fig. 2d) that aims to reduce the number of parameters, while preserving a certain semantic segmentation performance. DFANet encoder is constructed by first starting from the Xception model as a lightweight backbone, and then providing the high-level feature maps obtained from this backbone as an output to the next Xception backbone. This process, which is called sub-network aggregation, continues for three parallel Xception backbones. The goal of adding a fully-connected attention module to the tail of each backbone is to acquire the maximum receptive field. In addition to this sub-network aggregation, another module, namely the sub-stage aggregation, helps to ensure the combination of multi-scale information. By recovering the lost

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2022.3203145

spatial information caused by deep architectures, the aim of the sub-stage aggregation module is quite similar to that of the skip connections except that skip connections are not able to keep the large-scale object and edge information in very deep architectures. The sub-stage aggregation module is based on the fusion of different stages of the same depth in sub-networks. In other words, the output of the layer at a certain resolution from the previous backbone is contributing to the input of the corresponding layer of the next backbone.

The encoder, which is composed of the aggregation of three lightweight Xception backbones by means of sub-network and sub-stage modules, is followed by a slight decoder designed particularly to tackle real-time inference concerns. This decoder is conceived as an efficient feature upsampling module that is characterized by the convolution and bilinear upsampling layers. This simple decoder structure fuses the high-level features and the low-level features of the three backbones within itself. After the high-level features are bilinearly upsampled by a factor of 4, these are added to the low-level features. Finally, this total sum is upsampled by a factor of 4 and the final prediction is obtained.

Random Forest [68]: In order to compare the data-driven learning power of deep neural networks relative to traditional machine learning algorithms, the Random Forest (RF) method, which relies on hand-crafted features, is also introduced for segmentation performance evaluation. The RF method includes a set of decision trees in which each acts as a base classifier and depends on the independently sampled values of a random vector from the same distribution. As an ensemble learning algorithm, Random Forest builds many decision trees at the time of training and relies on the maximum voting of the decision trees in the forest. Its robustness to overfitting and good generalization ability are among the advantages of the RF algorithm in semantic segmentation [94], [95].

The hand-crafted features that we fed to the RF model are basic image features, specifically derivatives (up to second order) in three different scales. The number of decision trees in the Random Forest ensemble is selected as 20, whereas the minimum number of observations per tree leaf is set to 60.

C. Model Input

In this study, we are also interested in exploring the correlation of spectral bands and vegetation indices with tree segmentation performance. For that purpose, the benchmarked models are trained and tested with a set of input data with varying channel depth. We categorize the utilized model input into two main categories, three-channel and multispectral input. The first, the three-channel input, includes RGB, single vegetation indices (three channels being the same), the near-infrared (NIR) input (including separate three NIR sub-bands) and a three-channel vegetation index combination, including NDVI, ARVI, and SAVI indices (show below). Since this input group has a depth of three channels, DL models that comprise pre-trained encoders such as ResNet or Xception can be used in experiments.

The second group, multispectral input, includes image sets with a higher number of channel depths. The characteristics of the multispectral band input depend on the image set (i.e., the sensor) and are explained in detail below. As expected, experiments using this input group were only conducted for the U-Net and the SegNet DL models, which can accept any number of input channel depths. On the other hand, Random Forest experiments are applied to both model input groups.

Red-Green-Blue (RGB): As the name implies, these images consist of only Red, Green, and Blue channels. The difference between the reflected radiations in visible (i.e., RGB) and NIR wavelengths give more information about vegetation in particular. A large difference suggests that the pixels are more likely to belong to dense vegetation, such as forest area, while a small one is likely to indicate sparse vegetation, such as grassland [96]. On the other hand, for most commercial satellites (such as World-View-3), the spatial resolution of RGB channels is exceedingly higher than the original multispectral output of the sensor due to a resolution enhancement post-process applied to visible band output.

Vegetation Indices: Vegetation covers exhibit a unique spectral behaviour, by which they can be differentiated from other ground elements [64]. Commonly, chlorophyll concentration is responsible for absorbing the radiation in the Red band, while the leaf cellular structure is responsible for reflecting the radiation in the shorter NIR bands (700 to 750 nm). A deviation between the Red and NIR is observed in the spectral reflectance curve, providing deeper insight into the existence of vegetation. The key idea is to make use of this deviation to differentiate vegetation from bare soil based on the contrast between the spectral reflectance behaviours. Another spectral radiance difference measured between the Red and the Blue bands acts as a self-corrector by reducing the atmospheric scattering effects in the Red band, making it possible to define atmospherically resilient vegetation indices.

Current vegetation indices are commonly categorized in three fundamental families: mean vegetation indices, atmospherically resilient vegetation indices, and soil-adjusted vegetation indices [97]. For this reason, we utilized three widely-used indices from each family in our experiments. The key characteristics of these indices are delineated in the following:

Normalized Difference Vegetation Index (NDVI): is selected to represent the mean vegetation index category, for that it is one of the most widely used and well-known indices of this category. [98]. NDVI for the DSTL image set is defined by,

$$NDVI^D = \frac{NIR1 - Red}{NIR1 + Red} \quad , \tag{1}$$

where the WorldView-3 NIR1 band covers the range 0.77–0.90 μm and the Red band covers the range 0.63–0.69 μm.

In the case of the RIT-18 image set, NDVI is calculated as follows:

$$NDVI^R = \frac{NIR - Red}{NIR + Red} \quad , \tag{2}$$

where the RIT-18 NIR band covers the range 0.71–0.91 μm and the Red band covers the range 0.67–0.68 μm.

Atmospherically Resistant Vegetation Index (ARVI): ARVI is particularly designed to correct atmospheric effects. In order

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2022.3203145

to obtain such an atmospherically resilient VI, Blue or Green spectral bands are taken into account along with the Red and NIR bands. There are also some other related VIs that reduce atmospheric turbidity [99]. ARVI for the DSTL image set is given by,

$$ARVI^D = \frac{SWIR\ 1 - Red\ Edge - \gamma(Red\ Edge\ -\ Coastal)}{SWIR\ 1 + Red\ Edge - \gamma(Red\ Edge\ -\ Coastal)} \quad , \tag{3}$$

where the WorldView-3 SWIR 1 band covers the range 1.19–1.22 μm, Red Edge band covers 0.71–0.75 μm and the Coastal band covers the range 0.40–0.45 μm, and γ is taken as 1.0.

For the RIT-18 image set, ARVI calculation should be arranged according to the existing NIR bands due to the absence of the SWIR wavelengths. The following shows the calculation of ARVI for the RIT-18 image set:

$$ARVI^R = \frac{NIR - Red - \gamma(Red\ -\ Blue)}{NIR + Red - \gamma(Red\ -\ Blue)} \quad , \tag{4}$$

where the RIT-18 covers the NIR range 0.71–0.91 μm, the Red range 0.67–0.68 μm, and the Blue range 0.48–0.49 μm; γ is taken as 1.0.

Soil-Adjusted Vegetation Index (SAVI): Besides the two categories described above, there are many other VIs that have been proposed to reduce the soil background noise on NDVI by making use of a parameter called L, which incorporates the area density factor. One of these, SAVI enables the soil brightness effect to be assessed where the vegetation density is low in the area of interest [100]. For the DSTL image set, SAVI is given by,

$$SAVI^D = \frac{NIR1 - Red}{NIR1 + Red + L}(1 + L) \quad , \tag{5}$$

where L is taken as 0.5.

Using the NIR range of 0.71–0.91 m and a Red range of 0.67–0.68 m, SAVI can be formulated for the RIT-18 image set as follows:

$$SAVI^R = \frac{NIR - Red}{NIR + Red + L}(1 + L) \quad , \tag{6}$$

where L is taken as 0.5.

Mixed-Vegetation input (NDVIA+ARVI+SAVI): In addition to utilizing the aforementioned VIs separately as input to the benchmarked models, we propose a new three-channel input, called mixed-vegetation input, which uses three different VIs to test the combined effect of minimized soil brightness influences, corrected atmospheric scattering effects, and minimized topographic effects. The VIs selected from each of the categories are: NDVI from mean vegetation indices, ARVI from atmospherically resilient vegetation indices, and SAVI from soil-adjusted vegetation indices. The combined three-channel VI image is fed into the deep learning architectures just like an RGB input. A false-colour image of this fused input is depicted in Fig. 1.

Near Infrared input (NIRi): Both the DSTL and RIT-18 datasets utilized in our experiments are shot with sensors that output three NIR channels, details of which are provided in Table I. It can be seen from this table that, the spectral resolution of the RIT-18 dataset (TetraCam MicroMCA6) is finer than DSTL's sensor (World-View-3). On the other hand, World-View-3 NIR bands cover a larger spectrum. Hence we believe that testing these two 3-channel NIR input types in our experiments will provide us with an insight into the effect of NIR spectral resolution on tree segmentation.

Visible + Near Infrared input (VNIR): VNIR is one of the multispectral input groups utilized in our experiments. DSTL VNIR input includes 8 subbands (5 visible + 3 NIR), whereas RIT-18 includes 6 subbands (3 visible + 3 NIR) as a result of their individual sensory output (see Table I).

Visible + NIR + SWIR (VNIR+SWIR): VNIR+SWIR is the second multispectral input group utilized only for the DSTL image set in our experiments since SWIR bands are out of the RIT-18 sensor's spectral range. For this purpose two separate groups are constructed, the first being the VNIR+SWIR^2 input that includes 8 DSTL+VNIR subbands and the two shortest SWIR bands (i.e., SWIR-1 and SWIR-2). The second group is referred to as the VNIR+SWIR^8 and includes all of the original multispectral bands of the World-View-3 sensor in 16 channels.

IV. Experimental Results and Discussions

Before presenting the experimental results and the related discussions, we provide the details of the experimental setup and our evaluation methodology below.

A. Experimental Setup

A CUDA-enabled NVIDIA Quadro RTX 5000 GPU with 16GB memory is employed in the experiments. Other hardware configurations are: Intel(R) Xeon(R) Gold 6240R CPU @ 2.40GHz and 128GB RAM. The PyTorch [101] deep learning framework is used for training, validation, and testing on a Windows 10 operating system. In both of the image sets, a simple data pre-processing scheme is employed by dividing images into 224-by-224 image patches. The total number of image patches belonging to each image set can be seen in Table I.

The DSTL image set is split into training, validation, and test sets in which 20% of the data is reserved as a test set to perform cross-validation. The remaining 80% of the data is further partitioned into 72% training and 8% validation sets so that hyperparameter tuning can also be implemented. Five-fold cross-validation is applied to obtain a less biased estimate of the benchmarked methods. In order to improve segmentation performance, manual hyperparameter tuning is applied. All ED architectures are trained using the adaptive moment estimation (Adam) [102] algorithm. For the DSTL image set, the initial learning rate is chosen as 10^{-4} for U-Net and DFANet, while the initial value of $5*10^{-5}$ is used for the rest of the architectures, SegNet, DLinkNet and DeepLabv3+. For DLinkNet architecture, the learning rate is reduced by 9% in every 10 iteration steps, while for the others it is reduced by the same amount in every 5 iteration steps. Except for the SegNet, the mini-batch size is set to 8 and Xavier uniform is chosen for initialization. A Mini-batch size of 4 is chosen

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2022.3203145

for SegNet. By observing the learning curves, 70 epochs were found to be sufficient for convergence.

Due to the fact that test image labelling is not provided for the RIT-18 aerial imagery, it is possible to use the already partitioned training and validation sets [73], where the validation set is utilized as the test set in our experiments. For comparison, the same DSTL hyperparameters are also tuned for RIT-18. In order to handle the "black" (i.e., no data) regions in the image set (which are formed after image rectification), we did not utilize patches with more than 50% black region for training or testing. Thus, the training set is left with 814 image patches, while the test set contains 964. In the case of the RIT-18 image set, the same initial learning rate of 10^{-4} is used for SegNet, DLinkNet and DFANet architectures, while $5 * 10^{-5}$ is utilized for U-Net and DeepLabv3+. SegNet is the only model with its learning rate reduced by 9% in every 5 iteration steps, while other architectures have the same drop rate in every 10 iteration steps. The mini-batch size takes the value of 4 for SegNet and the DSTL image sets, while it is 8 for the rest. Xavier uniform is used as the initialization method and 100 epochs are evaluated with respect to convergence.

Evaluation metric: Intersection over Union (IoU or Jaccard Index) considers false alarms and missed values simultaneously by counting the total number of mislabeled pixels. The Jaccard Index is defined as,

$$Jaccard\ Index = \frac{TP}{TP + FP + FN} \quad , \tag{7}$$

where TP denotes the true positive pixels, FP the false positive pixels and FN the false-negative pixels. Jaccard Index is introduced as the evaluation metric since it is becoming the de-facto standard metric for semantic segmentation evaluation tasks in the literature. Since 2008, The Jaccard Index has been used in the PASCAL VOC challenge as the fundamental evaluation standard [103]. The average Jaccard Index (mJI) for all test images is provided in our results.

B. Results

Table II shows average Jaccard index values that are obtained from five-fold cross-validation results for the DSTL image set. As seen from the table, the best performance achieved is 0.571 (with a standard deviation of ±0.211 among all test images) by applying DLinkNet with the ResNet34 backbone to RGB images. The U-Net architecture applied to RGB images yields very close performance to that of DLinkNet, with 0.570 mJI. The closest performance of a non-RGB input type is again by the DLinkNet+ResNet34 architecture using the mixed vegetation indices, which yields an mJI of 0.561. We believe that the strength of the RGB input type on the DSTL image set is mainly because of the enhanced spatial resolution of the World-View-3's RGB channels. Although obtained from the original (i.e., unenhanced, lower) spatial resolution of the World-View-3 sensor, the mixed vegetation indices input type shows a competitive performance. These results show that for this sensor (i.e. with this spatial resolution and altitude) when fused within a three-channel signal, the hand-crafted VIs provide valuable low-level features to the DL architectures and hence perform as well as the enhanced RGB input type. We should note that the resolution enhancement post-process of RGB images from the World-View-3 sensor is an additional computation, which is carried out offline and may not always be supported for edge systems like satellites or UAVs.

On the other hand, even though the NIR reflectance information is quite important for assessing and distinguishing trees, the NIRi input type shows a low performance for all segmentation models for the DSTL image set. We believe that this is mainly due to two reasons: first, the aforementioned lower spatial resolution of the original NIR bands in the DSTL image set; and second, the coarse spectral resolution of the NIR subbands of the World-View-3 sensor, compared to RIT-18's performance with finer NIR bands as discussed below.

The segmentation performance results belonging to the RIT-18 image set are presented in Table III. There is a clear trend in NIRi results, with the highest values for most of the architectures. Judging by these superior mJI values, NIR reflectance shows a promising performance for tree segmentation. The peak value of 0.921 mJI is observed in the NIR reflectance band when DLinkNet with Resnet-34 encoder is applied. NIR wavelengths also work well for the U-Net architecture, with 0.885 mJI, giving evidence of the contribution from both high spatial (GSD 0.047 m) and high spectral resolution (3 channels between 485-685 nm).

To this extent, the segmentation performance of DLinkNet (Resnet-34) in the RIT-18 image set is consistent with those obtained in the DSTL image set. DLinkNet pre-trained with ResNet-34 reliably outperforms other ED architectures and is better able to preserve detailed spatial information. The combination of the VIs (NDVI+SAVI+ARVI) does not have a significant impact on the performance if an image set already has a very high spatial and spectral resolution for NIR reflectance; thus the NIR band alone can be a better choice for aerial remote sensing.

Another significant difference between the results of the DSTL and the RIT-18 image set experiments is the performance of the ARVI input type. Although performing comparatively well on the DSTL dataset, ARVI shows the lowest mJI values for the RIT-18 dataset experiments. As explained in the previous sections, atmospheric correction requires information from the SWIR band. Hence, the lower performance of ARVI on the RIT-18 image set is due to the lack of SWIR spectral bands in the image set's sensory output. When calculated properly using the required spectral bands, we observe a clearly improved performance in the hand-crafted VIs for tree segmentation.

The results of the experiments that utilize multispectral input types are provided in Table IV. Even though there are a number of band fusion modules applied by slightly modifying the networks [104], [105], the motivation in this study is to implement only the original architectures so only U-Net, SegNet, and random forest are used. Although tested for a limited number of architectures, the results are consistent with three-channel input type experiments. The lowest performance is obtained with the VNIR+SWIR^8 input type for the DSTL image set. We see that the effect of increasing the number of input channels by itself is not significant to

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2022.3203145

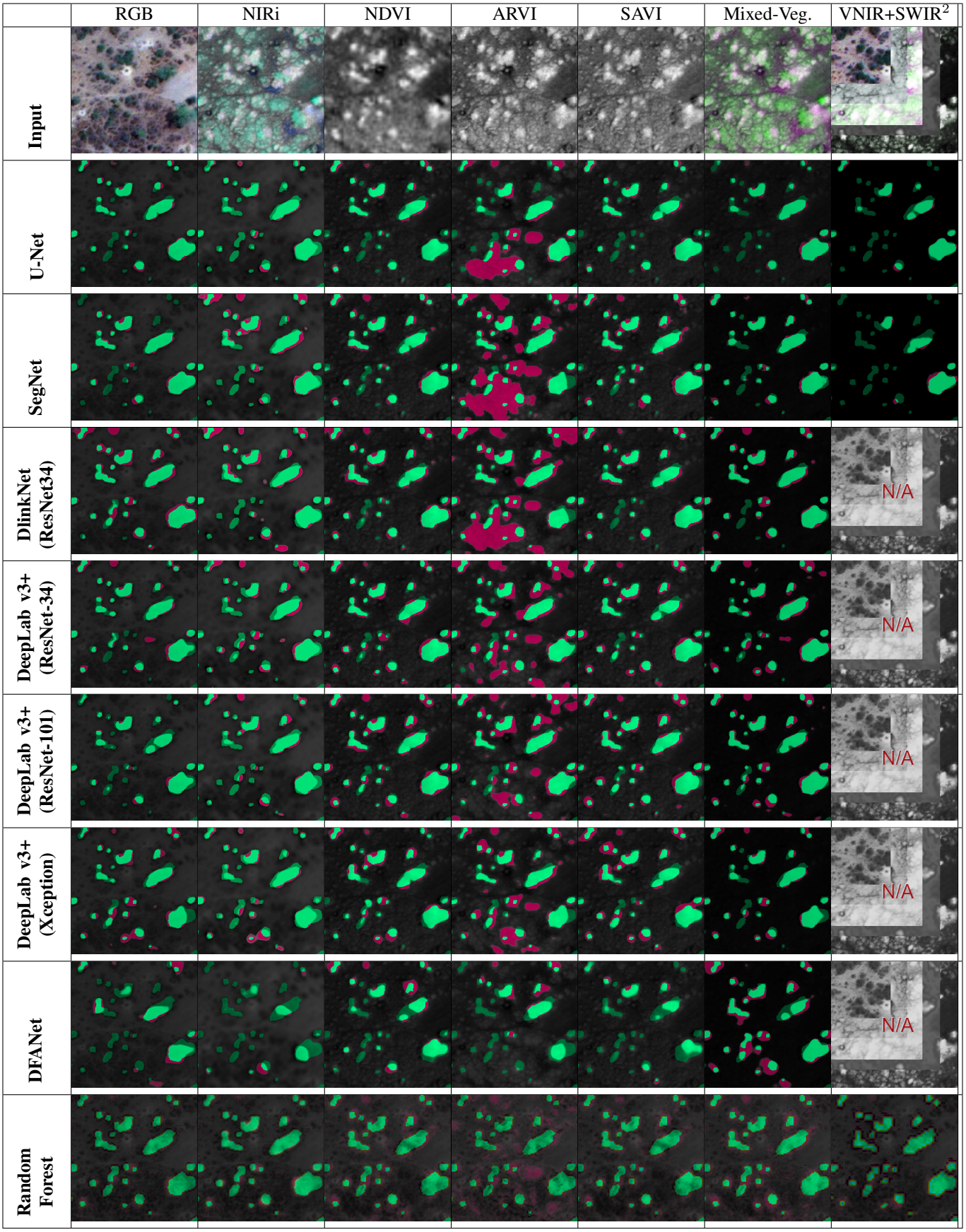


Fig. 3 Semantic segmentation results for all implemented models and utilized input types, depicted on a selected 224x224 pixels resolution region (approx. 70m×70m) from the **DSTL Image Set**. The first row consists of different input images; the rest are the semantic segmentation results. On semantic segmentation results, light green regions are the hits, dark green regions are the misses, and pink regions are the false alarms.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2022.3203145

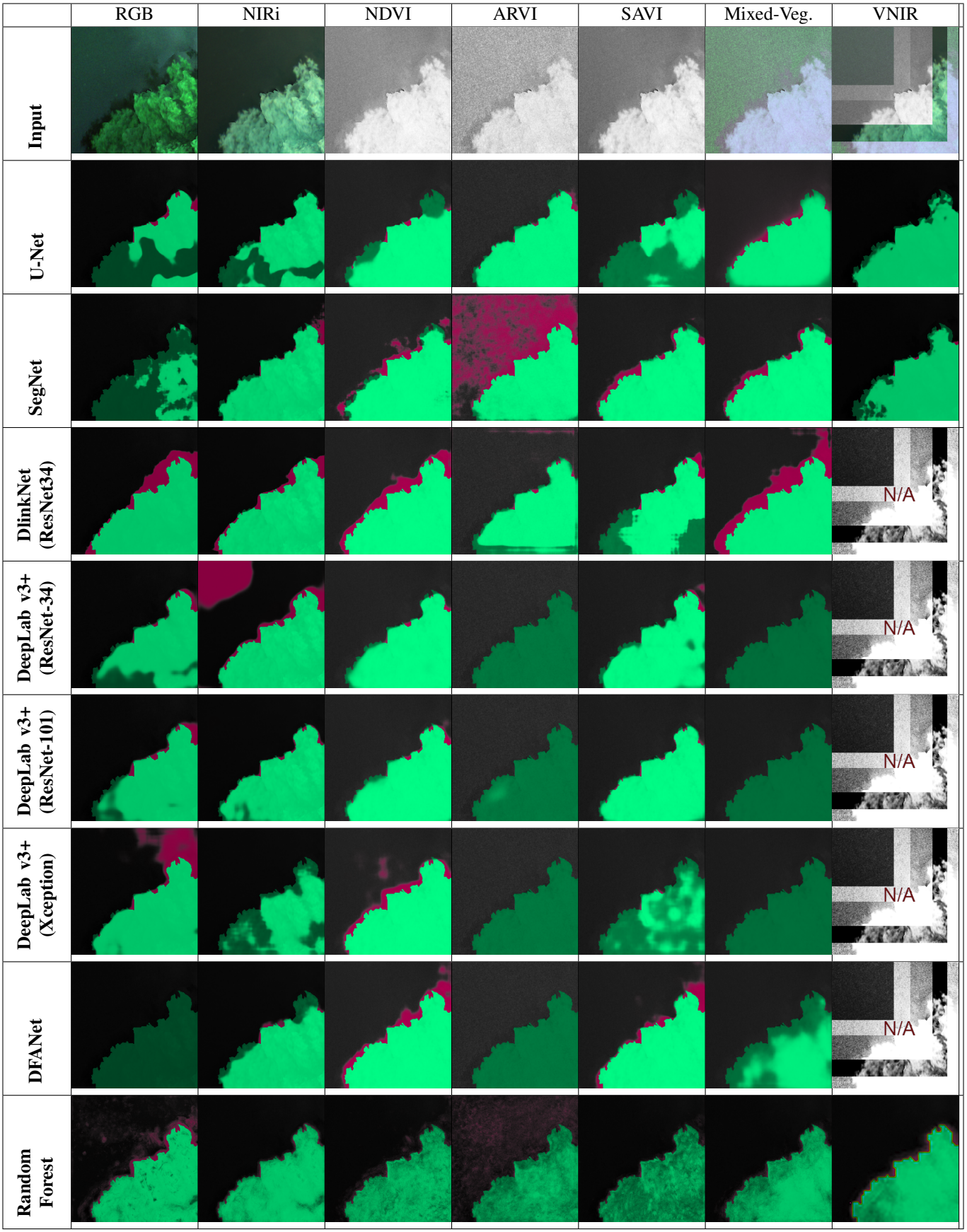


Fig. 4 Semantic segmentation results for all implemented models and utilized input types, depicted on a selected 224x224 pixels resolution region (apprx. 11m×11m) from the **RIT-18 Image Set**. The first row consists of different input images; the rest are the semantic segmentation results. On semantic segmentation results, light green regions are the hits, dark green regions are the misses, and pink regions are the false alarms.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2022.3203145

TABLE II
Results for the **DSTL Image Set** using the "Three-Channel Input" Group

	Jaccard Index					
	RGB	NIRi	NDVI	ARVI	SAVI	NDVI+ARVI+SAVI
U-Net	**0.570 ± 0.206**	0.493 ± 0.228	0.462 ± 0.253	0.493 ± 0.235	0.453 ± 0.259	0.524 ± 0.231
SegNet	0.492 ± 0.247	0.429 ± 0.259	0.415 ± 0.271	0.458 ± 0.246	0.416 ± 0.270	0.462 ± 0.246
DLinkNet (ResNet-34)	**0.571 ± 0.211**	0.526 ± 0.223	0.442 ± 0.262	0.485 ± 0.253	0.456 ± 0.257	**0.561 ± 0.226**
DeepLabv3+ (ResNet-34)	0.517 ± 0.212	0.504 ± 0.207	0.427 ± 0.244	0.435 ± 0.239	0.425 ± 0.246	0.505 ± 0.212
DeepLabv3+ (ResNet-101)	0.457 ± 0.217	0.466 ± 0.222	0.431 ± 0.252	0.440 ± 0.240	0.424 ± 0.244	0.475 ± 0.222
DeepLabv3+ (Xception)	0.481 ± 0.218	0.428 ± 0.232	0.313 ± 0.295	0.379 ± 0.248	0.309 ± 0.294	0.443 ± 0.244
DFANet	0.355 ± 0.239	0.350 ± 0.236	0.299 ± 0.264	0.309 ± 0.247	0.244 ± 0.263	0.313 ± 0.244
Random Forest	0.371 ± 0.183	0.300 ± 0.177	0.219 ± 0.176	0.232 ± 0.126	0.206 ± 0.172	0.316 ± 0.180

TABLE III
Results for the **RIT-18 Image Set** using the "Three-Channel Input" Group

	Jaccard Index					
	RGB	NIRi	NDVI	ARVI	SAVI	NDVI+ARVI+SAVI
U-Net	0.860 ± 0.285	**0.885 ± 0.263**	0.841 ± 0.306	0.758 ± 0.371	0.827 ± 0.315	0.715 ± 0.293
SegNet	0.852 ± 0.293	0.829 ± 0.330	0.835 ± 0.310	0.753 ± 0.340	0.828 ± 0.330	0.513 ± 0.327
DLinkNet (ResNet-34)	**0.893 ± 0.233**	**0.921 ± 0.208**	0.861 ± 0.289	0.748 ± 0.368	0.842 ± 0.317	0.803 ± 0.360
DeepLabv3+ (ResNet-34)	0.841 ± 0.311	0.872 ± 0.284	0.810 ± 0.356	0.684 ± 0.465	0.820 ± 0.336	0.682 ± 0.464
DeepLabv3+ (ResNet-101)	0.852 ± 0.294	0.869 ± 0.296	0.787 ± 0.373	0.685 ± 0.464	0.811 ± 0.348	0.684 ± 0.465
DeepLabv3+ (Xception)	0.858 ± 0.277	0.836 ± 0.308	0.815 ± 0.345	0.684 ± 0.465	0.811 ± 0.346	0.681 ± 0.465
DFANet	0.807 ± 0.342	0.837 ± 0.306	0.796 ± 0.322	0.710 ± 0.461	0.821 ± 0.314	0.707 ± 0.408
Random Forest	0.852 ± 0.237	0.866 ± 0.236	0.775 ± 0.288	0.732 ± 0.313	0.773 ± 0.287	0.821 ± 0.258

TABLE IV
Results for Both Sets using the "Multispectral Input" Group

	Jaccard Index			
	DSTL			**RIT-18**
	VNIR (5+3)	VNIR+SWIR^2 (5+3+2)	VNIR+SWIR^8 (5+3+8)	VNIR (3+3)
U-Net	0.546 ± 0.231	0.508 ± 0.237	0.538 ± 0.220	0.797 ± 0.351
SegNet	0.397 ± 0.268	0.427 ± 0.262	0.423 ± 0.263	0.841 ± 0.306
Random Forest	0.360 ± 0.184	0.358 ± 0.185	0.352 ± 0.185	0.862 ± 0.236

improve the segmentation performance; however, the optimum combination of these input channels is significant [106]–[108]. Literature already shows that because of an insufficient amount of labelled remote-sensing data, deep learning models are usually difficult to be trained in an end-to-end fashion using multispectral input [25]. It is one of our fundamental research questions in this paper to address the need for using VIs as hand-crafted features, instead of end-to-end training. In [109], it has already been shown that adding the input of extra-bands causes an inference to the network parameter learning and this, in turn, causes a degradation of the network performance. We believe that even if the input of all band data can increase the global information, those extra bands may be insufficient to identify detailed small tree objects in the scene [110].

Another point that cannot be ignored is most of these multiple spectral bands such as NIR and SWIR in the DSTL image set have low spatial resolutions compared to RGB, leading to a degradation in the overall segmentation performance [106], [111]. This is another indicator that supports the idea of using VIs as hand-crafted features when the spatial or spectral resolution of the visible and/or near-infrared bands is not sufficiently fine and the scale of the labelled image set is limited.

As obvious from Table III, the Jaccard index values of the random forest algorithm are almost comparable to those of the encoder-decoder architectures for the RIT-18 image set while this is not the case in Table II, i.e., for the DSTL image set. As already shown in [112], random forest can considerably improve the performance by increasing the number of spectral bands in high spatial resolution images. Moreover, the random forest algorithm not only offers significant performance for dealing with multi-dimensional complex data [113], [114], but also requires only slight parameter tuning [115]. Therefore, random forest is more likely to be robust to performance degradation when multiple bands are used. This is why the best result of the multispectral input in Table IV belongs to the random forest for the RIT-18 image set, where the spatial resolution is high.

The tree segmentation problem has many similarities to that of road segmentation in cases where multispectral remote sensing imagery is utilized. Occlusion is one of them, in which the object is partially or completely covered [116]. In the case of remote sensing imagery, the need for large receptive fields is crucial, since the input images are high resolution. The benefit of preserving detailed spatial information is additionally important, thus making the segmentation capable of acquiring complex geometries with highly discriminative feature information. We believe that these similarities, related to occlusions, high-resolution, and geometrical complexity, are the main reasons why a state-of-the-art road segmentation network such as DLinkNet is able to provide excellent results for tree segmentation as well.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2022.3203145

The similarities of our results can further be generalized to other road segmentation studies [80], [81], [117] as well. We observe that the performance of the DeepLabv3+ architecture employed with any of the state-of-the-art encoder backbones is falling short of effective occlusion handling. DeepLabv3+ models are highly optimized to the particular problem definition and so it is unsurprising that they show performance degradation in diverse remote sensing imagery. Hence, DeepLabv3+ cannot handle the occlusion problem and may perform below its potential due to the high demand for training samples [81].

Since our selection criterion in the ED architectures is mainly related to the decoder design, segmentation performance results should also be appraised with respect to the types of upsampling layers. U-Net and DLinkNet are expected to provide broadly similar results, as they both use the same upsampling layer of transposed convolution. Segmentation performance results confirm that the architectures designed with transposed convolution, i.e., U-Net and DLinkNet, achieve superior performance over other ED architectures. This behaviour is persistent regardless of the image sets and no matter what kind of input is fed to these networks. DLinkNet achieves a comparable and even better performance than U-Net by exploiting dilated convolution to expand the receptive field and by using pre-trained ResNet-34 as its encoder. The reason why SegNet fails to achieve superior performance is that the recovery of high-resolution spatial details using "index-based" skip connections is not sufficiently effective [81]. In DeepLabv3+, the implementation of bilinear interpolation upsampling may lead to a smooth segmentation so that the decoder is unable to recover pixel-wise tree predictions accurately [118], [119]. This is mainly due to the fact that the decoder in Deeplabv3+ is not suitable for processing high-resolution remote sensing imagery, where the tree pixels are small compared to the overall image [120].

The performance of the DFANet architecture should be further examined against that of the other architectures, since it is specifically tailored for real-time semantic segmentation. Just like the Deeplabv3+, DFANet also employs bilinear upsampling operation in the decoder part such that both have similar segmentation performances. However, as shown in Table II, DFANet Jaccard index values are even lower than those of the Deeplabv3+ due to the low spatial resolution of the DSTL image set. As a result of its lack of network depth and reduced number of parameters, DFANet cannot capture high-dimensional features like U-Net [121]. The DFANet experimental results are notable for their relatively high Jaccard index values for NDVI+SAVI+ARVI combination, which is consistent with our results obtained from other deep learning architectures. Moreover, the experimental results in Table III show that the segmentation performance of DFANet architecture is almost close to that of U-Net architecture, since the RIT-18 image set has high spatial resolution. We believe that, for the DFANet to be able to exhibit better semantic segmentation performance in high-resolution remote sensing imagery input, some improvements on the DFANet architecture should be done which allow the model to better fit to this kind of data [122]. Obviously, such improvements will also affect its real-time nature.

Although it is highly preferable to draw conclusions from averaged results such as those from Table II, III, and IV, it is worth noting that the single image results introduced in Fig. 3 and 4 depict explicit behaviour in which ARVI introduces significant false alarms. The pixels marked with pink are the false alarms, which can be seen to be quite dense in ARVI compared to other VIs. ARVI becomes sensitive to all kinds of green vegetation other than trees while reducing atmospheric scattering effects and hence presents false alarms by brightening the pixels belonging to other types of green vegetation, including small bushes and grasslands.

Tables II, III and IV show that the Random Forest algorithm achieves comparable segmentation performances, even though the results are not as good as those from deep learning-based ED architectures. Even the best of the RF algorithm results (RGB, with 0.371 mJI), shown in Table II for the DSTL image set, is still lower than all the DL-based results. On the other hand, the achieved results shown in Table III for the RIT-18 are quite high, with the highest being 0.866 mJI for NIR reflectance, indicating how well the RF algorithm can perform when spatial resolution is sufficiently high.

For all input types and utilized methods, results exhibit high standard deviations of Jaccard index values. This leads us to a conclusion that individual maximum or minimum performances of all input types and/or methods are statistically close to each other. We believe that the reason behind this fact is due to the tested images being significantly diverse in nature. In RIT-18 image set, some images cover only a car park, whereas some images only cover vegetation. In DSTL, some images cover only buildings, whereas some others cover only traffic and roads. Due to this extreme diversity, we believe that the high deviation in the individual results are justifiable. However, the deviation in Jaccard values is calculated based on the results that are obtained from thousands of images. The average Jaccard values (mJIs) obtained from this dataset are considerably different (reaching 0.3 mJI of difference between some input types and methods). We believe that, although there is a deviation of success related to the diverse nature of individual images, the mJI values obtained from thousands of samples are a true indicator of the capabilities of different input types and/or segmentation methods.

V. Conclusions and Future Directions

In this paper, we study different semantic segmentation models and different multispectral input combinations for the pixel-wise tree segmentation problem on remote sensing imagery. For this purpose, we utilize a set of comparative experiments where we benchmark the selected models and the given input types. The essence of the comparison for the segmentation models lies in the selection of ED architectures for tree segmentation. We specifically analyze the descriptive power of these models to overcome issues such as tree occlusion and geometric complexity. Different decoder designs, which make use of high-resolution information of remote sensing imagery and enable ED architectures to handle these problems, are explored. The comparative results procure

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2022.3203145

significant conclusions, the most important of which are summarized below.

1. The spatial resolution of the remote sensing imagery is the most important factor and has a greater influence on segmentation performance than spectral resolution and applied architectures. It is necessary to utilize a very high-resolution remote sensing image set: otherwise, it is not possible to reach considerable segmentation accuracy, even when a powerful, DL-based ED architecture is implemented. Therefore, an aerial image set is a better choice than satellite imagery, which is not only insufficient in terms of spatial resolution but also expensive.
2. The second most important factor is the spectral resolution of the NIR reflectance, for increasing the spectral discrimination between trees and other green vegetation. Although it is known that NIR reflectance information is significantly valuable for discriminating trees, the experiments demonstrate that if the spectral resolution of the NIR band is not sufficiently fine, sufficiently accurate tree segmentation cannot be obtained.
3. DLinkNet architecture consistently outperforms the compared semantic segmentation models, mainly for the cases including tree occlusion and grassland. Thanks to the decoder design with transposed convolution layers, further improved with dilated convolution and pre-trained ResNet-34, the large receptive field of the DLinkNet is best suited to high-resolution remote sensing images.
4. It can be advantageous to use a combination of vegetation indices in cases where the spatial resolution of the remote sensing image set is insufficient,such as the case we observe in our experiments with the DSTL image set. In TableII, we see that RGB and mixed-Veg input types perform similarly because the spatial resolution is not sufficiently high. This is not the case in Table III. Thus, we conclude that three factors should be taken into account when VIs are to be used for semantic segmentation of trees: (1) the input signal should include a combination of VIs of diverse characters, (2) the spatial resolution is relatively lower as in the case of satellite imagery, and (3) a DL-based model architecture with sufficiently descriptive strength should be utilized.
5. To assess trees in the presence of the most common types of aerosols such as smoke or sulfates, it is especially essential to utilize the SWIR band while calculating VIs that reduce atmospheric effects such as ARVI. Due to the sensitivity of the SWIR band to the liquid water content of the tree leaves, ARVI calculated with the SWIR band shows a better segmentation performance than the one that does not utilise SWIR reflectance.

Regarding the future directions for this study, our first plan is to focus on improving the backbone architecture. The pre-trained models such as the ResNet are trained on RGB data, and they can potentially downgrade the performance when fine-tuning on small remote sensing datasets due to the large domain gap. Because, as we saw in our preliminary experiments, it is not possible to obtain convergence for training, when we train some deep architectures (such as DeepLabv3+ or DLinkNet) from scratch with a limited scale of data. Hence, the only ideal option can be to use encoders pre-trained with satellite or aerial imagery with the same sensors, or in other words, the same multispectral bands trained with the same architectures. Such pre-trained encoders, to the best of our knowledge, do not publicly exist. However, there are efforts to implement domain adaptation to multispectral images obtained from different sensors, so that larger-scale experiments can be carried out [123]. Following this direction, a domain (i.e. sensor) independent multispectral backbone can be obtained.

As an alternative, converting the RGB backbones to a compatible model that can process multispectral data is also a viable direction. Additive Group Normalization method discussed in [104] is a recent example of this approach. By analyzing which bands contribute more to segmentation, they attempt to obtain deep-learning-based VI designs that will be compatible with three-channel backbones such as the ResNet. Following this path, we may be able to utilize multispectral band input to advanced architectures such as DLinkNet or DeepLabv3+, and achieve better performance.

Another promising direction, not only for this particular problem at hand but in deep learning is utilizing vision transformers (ViT) instead of convolution-based architectures. Following the very recent ViT design [124], computer vision problems are being attacked with transformer-based architectures. As of the time this paper is prepared, many transformed-based multispectral applications including segmentation [125], and many others [126], [127], are proposed, and will likely increase in near future.

REFERENCES

[1] T. M. Lenton, H. Held, E. Kriegler, J. W. Hall, W. Lucht, S. Rahmstorf, and H. J. Schellnhuber, "Tipping elements in the earth's climate system," *Proceedings of the National Academy of Sciences*, vol. 105, no. 6, pp. 1786–1793, 2008. [Online]. Available: https://www.pnas.org/content/105/6/1786

[2] L. Zhang, Z. Shao, J. Liu, and Q. Cheng, "Deep learning based retrieval of forest aboveground biomass from combined lidar and landsat 8 data," *Remote Sensing*, vol. 11, no. 12, p. 1459, 2019.

[3] D. J. Mulla, "Twenty five years of remote sensing in precision agriculture: Key advances and remaining knowledge gaps," *Biosystems engineering*, vol. 114, no. 4, pp. 358–371, 2013.

[4] S. Barati, B. Rayegani, M. Saati, A. Sharifi, and M. Nasri, "Comparison the accuracies of different spectral indices for estimation of vegetation cover fraction in sparse vegetated areas," *The Egyptian Journal of Remote Sensing and Space Science*, vol. 14, no. 1, pp. 49–56, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1110982311000147

[5] K. Herndon, R. Muench, E. Cherrington, and R. Griffin, "An assessment of surface water detection methods for water resource management in the nigerien sahel," *Sensors*, vol. 20, no. 2, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/2/431

[6] A. A. A. Alkhatib, "A review on forest fire detection techniques," *International Journal of Distributed Sensor Networks*, vol. 10, no. 3, p. 597368, 2014. [Online]. Available: https://doi.org/10.1155/2014/597368

[7] O. Ghorbanzadeh, K. Gholamnia, and P. Ghamisi, "The application of resu-net and obia for landslide detection from multi-temporal sentinel-2 images," *Big Earth Data*, pp. 1–26, 2022.

[8] X. Yuan, J. Shi, and L. Gu, "A review of deep learning methods for semantic segmentation of remote sensing imagery," *Expert Systems with Applications*, vol. 169, p. 114417, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417420310836

[9] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 117, pp. 11–28, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271616300144

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2022.3203145

[10] H. Mayer, "Automatic object extraction from aerial imagery—a survey focusing on buildings," *Computer Vision and Image Understanding*, vol. 74, no. 2, pp. 138–149, 1999. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077314299907506

[11] J. Mena, "State of the art on automatic road extraction for gis update: a novel classification," *Pattern Recognition Letters*, vol. 24, no. 16, pp. 3037–3058, 2003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865503001648

[12] F. A. Gougeon, "A crown-following approach to the automatic delineation of individual tree crowns in high spatial resolution aerial images," *Canadian journal of remote sensing*, vol. 21, no. 3, pp. 274–284, 1995.

[13] T. Brandtberg and F. Walter, "Automated delineation of individual tree crowns in high spatial resolution aerial images by multiple-scale analysis," *Machine Vision and Applications*, vol. 11, no. 2, pp. 64–73, 1998.

[14] P. Meyera, K. Staenzb, and K. Ittena, "Semi-automated procedures for tree species identification in high spatial resolution data from digitized colour infrared-aerial photography," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 51, no. 1, pp. 5–16, 1996.

[15] J. Wu, W. Yao, and P. Polewski, "Mapping individual tree species and vitality along urban road corridors with lidar and imaging sensors: Point density versus view perspective," *Remote Sensing*, vol. 10, no. 9, p. 1403, 2018.

[16] M. Immitzer, C. Atzberger, and T. Koukal, "Tree species classification with random forest using very high spatial resolution 8-band worldview-2 satellite data," *Remote sensing*, vol. 4, no. 9, pp. 2661–2693, 2012.

[17] V. Plakman, T. Janssen, N. Brouwer, and S. Veraverbeke, "Mapping species at an individual-tree scale in a temperate forest, using sentinel-2 images, airborne laser scanning data, and random forest classification," *Remote Sensing*, vol. 12, no. 22, p. 3710, 2020.

[18] L. Soleimannejad, S. Ullah, R. Abedi, M. Dees, and B. Koch, "Evaluating the potential of sentinel-2, landsat-8, and irs satellite images in tree species classification of hyrcanian forest of iran using random forest," *Journal of Sustainable Forestry*, vol. 38, no. 7, pp. 615–628, 2019.

[19] C. Sothe, C. De Almeida, M. Schimalski, L. La Rosa, J. Castro, R. Feitosa, M. Dalponte, C. Lima, V. Liesenberg, G. Miyoshi *et al.*, "Comparative performance of convolutional neural network, weighted and conventional support vector machine and random forest for classifying tree species using hyperspectral and photogrammetric data," *GIScience & Remote Sensing*, vol. 57, no. 3, pp. 369–394, 2020.

[20] R. Al-Ruzouq, A. Shanableh, B. A. Gibril, S. Al-Mansoori *et al.*, "Image segmentation parameter selection and ant colony optimization for date palm tree detection and mapping from very-high-spatial-resolution aerial imagery," *Remote Sensing*, vol. 10, no. 9, p. 1413, 2018.

[21] M. S. Colgan, C. A. Baldeck, J.-B. Féret, and G. P. Asner, "Mapping savanna tree species at ecosystem scales using support vector machine classification and brdf correction on airborne hyperspectral and lidar data," *Remote Sensing*, vol. 4, no. 11, pp. 3462–3480, 2012.

[22] K. Liyanage and B. M. Whitaker, "Satellite image classification using lc-ksvd sparse coding," in *2020 Intermountain Engineering, Technology and Computing (IETC)*. IEEE, 2020, pp. 1–6.

[23] L. Jin, S. Gao, Z. Li, and J. Tang, "Hand-crafted features or machine learnt features? together they improve rgb-d object recognition," in *2014 IEEE International Symposium on Multimedia*. IEEE, 2014, pp. 311–319.

[24] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 2, pp. 22–40, 2016.

[25] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, 2017.

[26] M. Wu, C. Zhang, J. Liu, L. Zhou, and X. Li, "Towards accurate high resolution satellite image semantic segmentation," *IEEE Access*, vol. 7, pp. 55 609–55 619, 2019.

[27] T. Hoeser and C. Kuenzer, "Object detection and image segmentation with deep learning on earth observation data: A review-part i: Evolution and recent trends," *Remote Sensing*, vol. 12, no. 10, p. 1667, 2020.

[28] T. Hoeser, F. Bachofer, and C. Kuenzer, "Object detection and image segmentation with deep learning on earth observation data: A review—part ii: Applications," *Remote Sensing*, vol. 12, no. 18, p. 3053, 2020.

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

[30] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.

[31] I. Ülkü and E. Akagündüz, "A survey on deep learning-based architectures for semantic segmentation on 2d images," *CoRR*, vol. abs/1912.10230, 2019. [Online]. Available: http://arxiv.org/abs/1912.10230

[32] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[33] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1520–1528.

[34] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[35] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[36] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[37] L. Zhou, C. Zhang, and M. Wu, "D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 182–186.

[38] H. Li, P. Xiong, H. Fan, and J. Sun, "Dfanet: Deep feature aggregation for real-time semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9522–9531.

[39] I. Ulku, P. Barmpoutis, T. Stathaki, and E. Akagunduz, "Comparison of single channel indices for U-Net based segmentation of vegetation in satellite images," in *Twelfth International Conference on Machine Vision (ICMV 2019)*, W. Osten and D. P. Nikolaev, Eds., vol. 11433, International Society for Optics and Photonics. SPIE, 2020, pp. 338 – 345. [Online]. Available: https://doi.org/10.1117/12.2556374

[40] B. Baheti, S. Innani, S. Gajre, and S. Talbar, "Semantic scene segmentation in unstructured environment with modified deeplabv3+," *Pattern Recognition Letters*, vol. 138, pp. 223–229, 2020.

[41] Y. Wang, B. Liang, M. Ding, and J. Li, "Dense semantic labeling with atrous spatial pyramid pooling and decoder for high-resolution remote sensing imagery," *Remote Sensing*, vol. 11, no. 1, p. 20, 2019.

[42] S. Du, S. Du, B. Liu, and X. Zhang, "Incorporating deeplabv3+ and object-based image analysis for semantic segmentation of very high resolution remote sensing images," *International Journal of Digital Earth*, pp. 1–22, 2020.

[43] S. Hartling, V. Sagan, P. Sidike, M. Maimaitijiang, and J. Carron, "Urban tree species classification using a worldview-2/3 and lidar data fusion approach and deep learning," *Sensors*, vol. 19, no. 6, p. 1284, 2019.

[44] K. A. Korznikov, D. E. Kislov, J. Altman, J. Doležal, A. S. Vozmishcheva, and P. V. Krestov, "Using u-net-like deep convolutional neural networks for precise tree recognition in very high resolution rgb (red, green, blue) satellite images," *Forests*, vol. 12, no. 1, p. 66, 2021.

[45] L. Ding, H. Tang, and L. Bruzzone, "Lanet: Local attention embedding to improve the semantic segmentation of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[46] S. Solberg, E. Naesset, and O. M. Bollandsas, "Single tree segmentation using airborne laser scanner data in a structurally heterogeneous spruce forest," *Photogrammetric Engineering & Remote Sensing*, vol. 72, no. 12, pp. 1369–1378, 2006.

[47] Q. Li, P. Yuan, X. Liu, and H. Zhou, "Street tree segmentation from mobile laser scanning data," *International Journal of Remote Sensing*, vol. 41, no. 18, pp. 7145–7162, 2020.

[48] W. Yan, H. Guan, L. Cao, Y. Yu, C. Li, and J. Lu, "A self-adaptive mean shift tree-segmentation method using uav lidar data," *Remote Sensing*, vol. 12, no. 3, p. 515, 2020.

[49] J. Yang, Z. Kang, S. Cheng, Z. Yang, and P. H. Akwensi, "An individual tree segmentation method based on watershed algorithm

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2022.3203145

and three-dimensional spatial distribution analysis from airborne lidar point clouds," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 1055–1067, 2020.

[50] C. Zhang, Y. Zhou, and F. Qiu, "Individual tree segmentation from lidar point clouds for urban forest inventory," *Remote Sensing*, vol. 7, no. 6, pp. 7892–7913, 2015.

[51] H. Hamraz, M. A. Contreras, and J. Zhang, "A robust approach for tree segmentation in deciduous forests using small-footprint airborne lidar data," *International Journal of Applied Earth Observation and Geoinformation*, vol. 52, pp. 532–541, 2016.

[52] F. Pirotti, M. Kobal, and J. Roussel, "A comparison of tree segmentation methods using very high density airborne laser scanner data." *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, vol. 42, 2017.

[53] S. Dersch, M. Heurich, N. Krueger, and P. Krzystek, "Combining graph-cut clustering with object-based stem detection for tree segmentation in highly dense airborne lidar point clouds," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 172, pp. 207–222, 2021.

[54] L. Comesaña-Cebral, J. Martínez-Sánchez, H. Lorenzo, and P. Arias, "Individual tree segmentation method based on mobile backpack lidar point clouds," *Sensors*, vol. 21, no. 18, p. 6007, 2021.

[55] K. Itakura, S. Miyatani, and F. Hosoi, "Estimating tree structural parameters via automatic tree segmentation from lidar point cloud data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 555–564, 2021.

[56] J. Hyyppa, O. Kelle, M. Lehikoinen, and M. Inkinen, "A segmentation-based method to retrieve stem volume estimates from 3-d tree height models produced by laser scanners," *IEEE Transactions on geoscience and remote sensing*, vol. 39, no. 5, pp. 969–975, 2001.

[57] J. B. Roerdink and A. Meijster, "The watershed transform: Definitions, algorithms and parallelization strategies," *Fundamenta informaticae*, vol. 41, no. 1, 2, pp. 187–228, 2000.

[58] E. Ayrey, S. Fraver, J. A. Kershaw Jr, L. S. Kenefic, D. Hayes, A. R. Weiskittel, and B. E. Roth, "Layer stacking: A novel algorithm for individual forest tree segmentation from lidar point clouds," *Canadian Journal of Remote Sensing*, vol. 43, no. 1, pp. 16–27, 2017.

[59] C. Dechesne, C. Mallet, A. Le Bris, and V. Gouet-Brunet, "Semantic segmentation of forest stands of pure species combining airborne lidar data and very high resolution multispectral imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 126, pp. 129–145, 2017.

[60] P. Krzystek, A. Serebryanyk, C. Schnörr, J. Červenka, and M. Heurich, "Large-scale mapping of tree species and dead trees in šumava national park and bavarian forest national park using lidar and multispectral imagery," *Remote Sensing*, vol. 12, no. 4, p. 661, 2020.

[61] D. Pulido, J. Salas, M. Rös, K. Puettmann, and S. Karaman, "Assessment of tree detection methods in multispectral aerial images," *Remote Sensing*, vol. 12, no. 15, p. 2379, 2020.

[62] S. T. Digumarti, L. M. Schmid, G. M. Rizzi, J. Nieto, R. Siegwart, P. Beardsley, and C. Cadena, "An approach for semantic segmentation of tree-like vegetation," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 1801–1807.

[63] M. Brandt, C. J. Tucker, A. Kariryaa, K. Rasmussen, C. Abel, J. Small, J. Chave, L. V. Rasmussen, P. Hiernaux, A. A. Diouf *et al.*, "An unexpectedly large count of trees in the west african sahara and sahel," *Nature*, vol. 587, no. 7832, pp. 78–82, 2020.

[64] A. Bannari, D. Morin, F. Bonn, and A. Huete, "A review of vegetation indices," *Remote sensing reviews*, vol. 13, no. 1-2, pp. 95–120, 1995.

[65] A. R. Huete, "Vegetation indices, remote sensing and forest monitoring," *Geography Compass*, vol. 6, no. 9, pp. 513–532, 2012.

[66] D. R. Peddle, S. Brunke, and F. G. Hall, "A comparison of spectral mixture analysis and ten vegetation indices for estimating boreal forest biophysical information from airborne data," *Canadian Journal of Remote Sensing*, vol. 27, no. 6, pp. 627–635, 2001.

[67] S. P. Mohanty, J. Czakon, K. A. Kaczmarek, A. Pyskir, P. Tarasiewicz, S. Kunwar, J. Rohrbach, D. Luo, M. Prasad, S. Fleer, J. P. Göpfert, A. Tandon, G. Mollard, N. Rayaprolu, M. Salathe, and M. Schilling, "Deep learning for understanding satellite imagery: An experimental survey," *Frontiers in Artificial Intelligence*, vol. 3, p. 85, 2020. [Online]. Available: https://www.frontiersin.org/article/10.3389/frai.2020.534696

[68] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[69] S. Madonsela, M. A. Cho, R. Mathieu, O. Mutanga, A. Ramoelo, Ż. Kaszta, R. Van De Kerchove, and E. Wolff, "Multi-phenology worldview-2 imagery improves remote sensing of savannah tree species," *International journal of applied earth observation and geoinformation*, vol. 58, pp. 65–73, 2017.

[70] A. Darshik, A. Dev, M. Bharath, B. A. Nair, and G. Gopakumar, "Semantic segmentation of spectral images: a comparative study using fcn8s and u-net on rit-18 dataset," in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. IEEE, 2020, pp. 1–6.

[71] H. Hao, S. Baireddy, K. LaTourette, L. Konz, M. Chan, M. L. Comer, and E. J. Delp, "Improving building segmentation using uncertainty modeling and metadata injection," in *Proceedings of the 29th International Conference on Advances in Geographic Information Systems*, 2021, pp. 117–120.

[72] A. Song and Y. Kim, "Semantic segmentation of remote-sensing imagery using heterogeneous big data: International society for photogrammetry and remote sensing potsdam and cityscape datasets," *ISPRS International Journal of Geo-Information*, vol. 9, no. 10, p. 601, 2020.

[73] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning," *ISPRS journal of photogrammetry and remote sensing*, vol. 145, pp. 60–77, 2018.

[74] L. Mou, Y. Hua, and X. X. Zhu, "A relation-augmented fully convolutional network for semantic segmentation in aerial scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 416–12 425.

[75] R. Kemker, C. Salvaggio, and C. Kanan, "High-resolution multispectral dataset for semantic segmentation," *arXiv preprint arXiv:1703.01918*, 2017.

[76] https://www.kaggle.com/c/dstl-satellite-imageryfeature-detection.

[77] http://www.satimagingcorp.com/.

[78] Z. Wojna, V. Ferrari, S. Guadarrama, N. Silberman, L.-C. Chen, A. Fathi, and J. Uijlings, "The devil is in the decoder: Classification, regression and gans," *International Journal of Computer Vision*, vol. 127, no. 11, pp. 1694–1706, 2019.

[79] http://host.robots.ox.ac.uk:8080/leaderboard/displaylb_main.php?challengeid=11&compid=6.

[80] Z. Tian, T. He, C. Shen, and Y. Yan, "Decoders matter for semantic segmentation: Data-dependent decoding enables flexible feature aggregation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3126–3135.

[81] D. Lobo Torres, R. Queiroz Feitosa, P. Nigri Happ, L. Elena Cue La Rosa, J. Marcato Junior, J. Martins, P. Ola Bressan, W. N. Gonçalves, and V. Liesenberg, "Applying fully convolutional architectures for semantic segmentation of a single tree species in urban environment on high resolution uav optical imagery," *Sensors*, vol. 20, no. 2, p. 563, 2020.

[82] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[83] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.

[84] ——, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[85] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[86] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[87] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[88] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[89] A. Chaurasia and E. Culurciello, "Linknet: Exploiting encoder representations for efficient semantic segmentation," in *2017 IEEE Visual Communications and Image Processing (VCIP)*. IEEE, 2017, pp. 1–4.

[90] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," *arXiv preprint arXiv:1606.02147*, 2016.

[91] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, and H. Hajishirzi, "Espnet: Efficient spatial pyramid of dilated convolutions for semantic

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2022.3203145

segmentation," in *Proceedings of the european conference on computer vision (ECCV)*, 2018, pp. 552–568.

[92] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "Icnet for real-time semantic segmentation on high-resolution images," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 405–420.

[93] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 325–341.

[94] J. Su, D. Yi, B. Su, Z. Mi, C. Liu, X. Hu, X. Xu, L. Guo, and W.-H. Chen, "Aerial visual perception in smart farming: Field study of wheat yellow rust monitoring," *IEEE transactions on industrial informatics*, vol. 17, no. 3, pp. 2242–2249, 2020.

[95] R. M Rustowicz, R. Cheong, L. Wang, S. Ermon, M. Burke, and D. Lobell, "Semantic segmentation of crop type in africa: A novel dataset and analysis of deep learning methods," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 75–82.

[96] https://earthobservatory.nasa.gov/features/MeasuringVegetation/measuring_vegetation_2.php.

[97] A. Azadeh, P. Dimitrios, and S. Peter, "Forest canopy density assessment using different approaches–review," *Journal of forest science*, vol. 63, no. 3, pp. 107–116, 2017.

[98] J. Rouse, R. H. Haas, J. A. Schell, D. W. Deering *et al.*, "Monitoring vegetation systems in the great plains with erts," *NASA special publication*, vol. 351, no. 1974, p. 309, 1974.

[99] Y. J. Kaufman and D. Tanre, "Atmospherically resistant vegetation index (arvi) for eos-modis," *IEEE transactions on Geoscience and Remote Sensing*, vol. 30, no. 2, pp. 261–270, 1992.

[100] A. R. Huete, "A soil-adjusted vegetation index (savi)," *Remote sensing of environment*, vol. 25, no. 3, pp. 295–309, 1988.

[101] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, pp. 8026–8037, 2019.

[102] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[103] G. Csurka, D. Larlus, F. Perronnin, and F. Meylan, "What is a good evaluation measure for semantic segmentation?" *IEEE PAMI*, vol. 26, no. 1, 2004.

[104] H. Sheng, X. Chen, J. Su, R. Rajagopal, and A. Ng, "Effective data fusion with generalized vegetation index: Evidence from land cover segmentation in agriculture," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 60–61.

[105] K. Yuan, X. Zhuang, G. Schaefer, J. Feng, L. Guan, and H. Fang, "Deep-learning-based multispectral satellite image segmentation for water body detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 7422–7434, 2021.

[106] M. A. E. Bhuiyan, C. Witharana, A. K. Liljedahl, B. M. Jones, R. Daanen, H. E. Epstein, K. Kent, C. G. Griffin, and A. Agnew, "Understanding the effects of optimal combination of spectral bands on deep learning model predictions: a case study based on permafrost tundra landform mapping using high resolution multispectral satellite imagery," *Journal of Imaging*, vol. 6, no. 9, p. 97, 2020.

[107] Y. Cai, H. Huang, K. Wang, C. Zhang, L. Fan, and F. Guo, "Selecting optimal combination of data channels for semantic segmentation in city information modelling (cim)," *Remote Sensing*, vol. 13, no. 7, p. 1367, 2021.

[108] H. Grybas and R. G. Congalton, "A comparison of multi-temporal rgb and multispectral uas imagery for tree species classification in heterogeneous new hampshire forests," *Remote Sensing*, vol. 13, no. 13, p. 2631, 2021.

[109] Z. Wang, P. Yang, H. Liang, C. Zheng, J. Yin, Y. Tian, and W. Cui, "Semantic segmentation and analysis on sensitive parameters of forest fire smoke using smoke-unet and landsat-8 imagery," *Remote Sensing*, vol. 14, no. 1, p. 45, 2022.

[110] W. Kang, Y. Xiang, F. Wang, and H. You, "Do-net: Dual-output network for land cover classification from optical remote sensing images," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.

[111] T. Yang, S. Jiang, Z. Hong, Y. Zhang, Y. Han, R. Zhou, J. Wang, S. Yang, X. Tong, and T.-y. Kuc, "Sea-land segmentation using deep learning techniques for landsat-8 oli imagery," *Marine Geodesy*, vol. 43, no. 2, pp. 105–133, 2020.

[112] X. Yu, D. Lu, X. Jiang, G. Li, Y. Chen, D. Li, and E. Chen, "Examining the roles of spectral, spatial, and topographic features in improving land-cover and forest classifications in a subtropical region," *Remote Sensing*, vol. 12, no. 18, p. 2907, 2020.

[113] M. W. Ahmad, M. Mourshed, and Y. Rezgui, "Trees vs neurons: Comparison between random forest and ann for high-resolution prediction of building energy consumption," *Energy and buildings*, vol. 147, pp. 77–89, 2017.

[114] M. Sheykhmousa, M. Mahdianpari, H. Ghanbari, F. Mohammadimanesh, P. Ghamisi, and S. Homayouni, "Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 6308–6325, 2020.

[115] J. Xia, P. Ghamisi, N. Yokoya, and A. Iwasaki, "Random forest ensembles and extended multiextinction profiles for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 1, pp. 202–216, 2017.

[116] S. Wang, H. Yang, Q. Wu, Z. Zheng, Y. Wu, and J. Li, "An improved method for road extraction from high-resolution remote-sensing images that enhances boundary information," *Sensors*, vol. 20, no. 7, p. 2064, 2020.

[117] M. Zhou, H. Sui, S. Chen, J. Wang, and X. Chen, "Bt-roadnet: A boundary and topologically-aware neural network for road extraction from high-resolution remote sensing imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 168, pp. 288–306, 2020.

[118] B. Zhang, Z. Tian, C. Shen *et al.*, "Dynamic neural representational decoders for high-resolution semantic segmentation," in *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[119] C. Shen, L. Liu, L. Zhu, J. Kang, N. Wang, and L. Shao, "High-throughput in situ root image segmentation based on the improved deeplabv3+ method," *Frontiers in Plant Science*, vol. 11, 2020.

[120] Y. Bao and Y. Zheng, "Based on the improved deeplabv3+ remote sensing image semantic segmentation algorithm," in *2021 4th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)*. IEEE, 2021, pp. 717–720.

[121] Y. Zhang, H. Ren, W. Yang, Y. Wang, K. Ye, and C.-Z. Xu, "The strong substructure and feature attention mechanism for image semantic segmentation," *Concurrency and Computation: Practice and Experience*, vol. 34, no. 12, p. e5920, 2022.

[122] Z. Wang, J. Guo, W. Huang, and S. Zhang, "High-resolution remote sensing image semantic segmentation based on a deep feature aggregation network," *Measurement Science and Technology*, vol. 32, no. 9, p. 095002, 2021.

[123] D. Lunga, H. L. Yang, A. Reith, J. Weaver, J. Yuan, and B. Bhaduri, "Domain-adapted convolutional networks for satellite image classification: A large-scale interactive learning workflow," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 3, pp. 962–977, 2018.

[124] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[125] Z. Xu, W. Zhang, T. Zhang, Z. Yang, and J. Li, "Efficient transformer for remote sensing image segmentation," *Remote Sensing*, vol. 13, no. 18, 2021. [Online]. Available: https://www.mdpi.com/2072-4292/13/18/3585

[126] Y. Bazi, L. Bashmal, M. Alrahhal, R. Dayil, and N. Ajlan, "Vision transformers for remote sensing image classification," *Remote Sensing*, vol. 13, p. 516, 02 2021.

[127] L. Bashmal, Y. Bazi, and M. A. Rahhal, "Deep vision transformers for remote sensing scene classification," in *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, 2021, pp. 2815–2818.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. This is the author's version which has not been fully e
content may change prior to final publication. Citation information: DOI 10.1109/JSTARS.2022.3203145



Irem Ulku received B.Sc. degrees in both Electronics and Communication Engineering and in Industrial Engineering (as valedictorian) from Çankaya University (Ankara, Turkey) in 2009 and 2010, respectively, followed by an M.Sc. degree in Electrical and Electronics Engineering from the Middle East Technical University (Ankara, Turkey) in 2013. She then received her PhD degree in Electronics and Communication Engineering from Çankaya University in 2017, where she also worked as an instructor in the same Department between 2017 to 2019 and 2020 to 2021. In 2019, she was a research associate at Imperial College London, United Kingdom. She is currently an assistant professor at the Department of Computer Engineering, Ankara University (Turkey). Her research interests include multispectral image processing and deep learning-based semantic segmentation.



Erdem Akagunduz is currently an associate professor at the Graduate School of Informatics, Middle East Technical University (METU), in Ankara, Turkey. He received his B.Sc., M.Sc., and PhD degrees in Electrical and Electronics Engineering from METU in 2001, 2004, and 2011, respectively. From 2001 to 2008, he was a research assistant with the METU Computer Vision and Intelligent Systems Laboratory, and from 2009 to 2016 he worked as a Computer Vision Scientist with ASELSAN Inc. (Ankara, Turkey). He then became a research associate at the University of York (United Kingdom) in 2016. Before joining back to METU, he worked as an assistant professor at Çankaya University, Electrical and Electronics Engineering Department. His research interests include infrared computer vision, object/target/scene recognition and deep learning.



Pedram Ghamisi (S'12, M'15, SM'18) graduated with a Ph.D. in electrical and computer engineering at the University of Iceland in 2015. He works as (1) the head of the machine learning group at Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Germany, and (2) research professor and group leader of AI4RS at the Institute of Advanced Research in Artificial Intelligence (IARAI), Austria. He is a co-founder of VasoGnosis Inc. with two branches in San Jose and Milwaukee, USA. He was the co-chair of the IEEE Image Analysis and Data Fusion Committee (IEEE IADF) between 2019 and 2021. Dr. Ghamisi was a recipient of the IEEE Mikio Takagi Prize for winning the Student Paper Competition at IEEE International Geoscience and Remote Sensing Symposium (IGARSS) in 2013, the first prize of the data fusion contest organized by the IEEE IADF in 2017, the Best Reviewer Prize of IEEE Geoscience and Remote Sensing Letters in 2017, and the IEEE Geoscience and Remote Sensing Society 2020 Highest-Impact Paper Award. His research interests include interdisciplinary research on machine (deep) learning, image and signal processing, and multisensor data fusion. He is an associate editor of IEEE JSTARS and IEEE GRSL. For detailed info, please see www.pedramghamisi.com/ and www.ai4rs.com.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/