ISPRS Journal of Photogrammetry and Remote Sensing 169 (2020) 280–291

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Identifying and mapping individual plants in a highly diverse high-elevation ecosystem using UAV imagery and deep learning

Ce Zhang [a,b,*], Peter M. Atkinson [c,*], Charles George [d], Zhaofei Wen [e], Mauricio Diazgranados [f], France Gerard [d]

[a] Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, UK
[b] UK Centre for Ecology & Hydrology, Library Avenue, Bailrigg, Lancaster LA1 4AP, UK
[c] Faculty of Science and Technology, Lancaster University, Lancaster LA1 4YR, UK
[d] UK Centre for Ecology & Hydrology, Maclean Building, Benson Lane, Wallingford OX10 8BB, UK
[e] Key Laboratory of Reservoir Aquatic Environment, Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing 400714, China
[f] Royal Botanic Gardens, Kew, Ardingly, West Sussex RH17 6TN, UK

ARTICLE INFO

Keywords:
Multi-scale deep learning
Residual U-Net
Scale sequence
Semantic segmentation
Páramos

ABSTRACT

The identification and counting of plant individuals is essential for environmental monitoring. UAV based imagery offer ultra-fine spatial resolution and flexibility in data acquisition, and so provide a great opportunity to enhance current plant and *in-situ* field surveying. However, accurate mapping of individual plants from UAV imagery remains challenging, given the great variation in the sizes and geometries of individual plants and in their distribution. This is true even for deep learning based semantic segmentation and classification methods. In this research, a novel Scale Sequence Residual U-Net (SS Res U-Net) deep learning method was proposed, which integrates a set of Residual U-Nets with a sequence of input scales that can be derived automatically. The SS Res U-Net classifies individual plants by continuously increasing the patch scale, with features learned at small scales passing gradually to larger scales, thus, achieving multi-scale information fusion while retaining fine spatial details of interest. The SS Res U-Net was tested to identify and map frailejones (all plant species of the subtribe Espeletiinae), the dominant plants in one of the world's most biodiverse high-elevation ecosystems (i.e. the páramos) from UAV imagery. Results demonstrate that the SS Res U-Net has the ability to self-adapt to variation in objects, and consistently achieved the highest classification accuracy (91.67% on average) compared with four state-of-the-art benchmark approaches. In addition, SS Res U-Net produced the best performances in terms of both robustness to training sample size reduction and computational efficiency compared with the benchmarks. Thus, SS Res U-Net shows great promise for solving remotely sensed semantic segmentation and classification tasks, and more general machine intelligence. The prospective implementation of this method to identify and map frailejones in the páramos will benefit immensely the monitoring of their populations for conservation assessments and management, among many other applications.

1. Introduction

The identification and counting of plant individuals is essential for environmental monitoring, whether this is in the context of plant population, habitat condition or ecosystem services assessment, crop yield estimation, invasive species or weed control action, or climate or disturbance impact assessment. While the most accurate approach to identifying individual plants and establishing their density is still mainly through *in situ* field surveying, the proliferation of technology delivering high-quality, high-definition imagery and recent developments in artificial intelligence and digital image processing, has opened up opportunities for automation.

Unmanned aerial vehicles (UAV) or drones, as most recent advances in sensors and platforms, have pushed the boundaries of spatial resolution from metre to sub-metre and towards state-of-the-art ultra-fine centimetre level (Aasen et al., 2018). Data from UAV-based imaging systems, often acquired at relatively low cost, can capture individual plants in extremely fine spatial detail, making them particularly attractive for mapping ecological species (Baena et al., 2017; Woellner and Wagner, 2019). However, UAV data acquisition, in terms of mission

* Corresponding authors.
E-mail addresses: c.zhang9@lancaster.ac.uk (C. Zhang), pma@lancaster.ac.uk (P.M. Atkinson).

https://doi.org/10.1016/j.isprsjprs.2020.09.025
Received 20 May 2020; Received in revised form 30 July 2020; Accepted 30 September 2020
Available online 8 October 2020

planning, navigation and orientation, is still in its infancy (Colomina and Molina, 2014) and images captured across and between sites are often at different spatial resolutions, due to the variation of in- and between-flight altitudes and platform orientations. Information extraction from these fine spatial resolution UAV images is challenging with exceptional levels of detail and complex shading patterns. Additionally, the current procedures are inefficient and lack automation, involving primarily aerial photo interpretation methods that are labour-intensive and time-consuming (Milas et al., 2017). Only the development of highly efficient and automated techniques to extract ecological information from ultra-fine spatial resolution UAV imagery will enable the operational use of UAV-based remote sensing in environmental research and monitoring.

Ultra-fine spatial resolution UAV imagery is only available as colour (visible spectrum) and near infrared photography (Hamylton et al., 2020). Therefore, to fully exploit the unprecedented details within the images, automated techniques for information retrieval have to move beyond the use of spectral features only, and include spatial characteristics such as spatial texture, shape, context and orientation (Kraaijenbrink et al., 2016). Traditional pixel-based classification methods, such as support vector machine, random forest, etc., were designed to only use multi-spectral information of a pixel (Zhang et al., 2018b). As a result, when applied to drone imagery, the rich *spatial* information presented in the imagery is not considered, leading to low classification accuracy with salt-and-pepper noise effects (Hamylton et al., 2020). Object-based image analysis (OBIA) has been adopted widely over the last decade as an alternative technique to enable the inclusion of spatial information in image classification (Blaschke et al., 2014). However, the spatial features captured in OBIA are based on information primarily from the pixels within each object (e.g. texture, pattern and shape). There is little consideration of the wider spatial and spectral context and the complexities present in the available drone imagery. As a result, identification of complex objects using OBIA approaches is limited in terms of accuracy.

Recently, deep learning techniques, and deep convolutional neural networks (CNN) in particular, have gained enormous interest in computer vision and pattern recognition, with state-of-the-art breakthroughs in image analysis and feature representations (Krizhevsky et al., 2012). The major advantage of CNN-based approaches in comparison with traditional methods is their ability to learn the most robust object characteristics through deep networks, building on the identification of features such as object edges, textures and parts, enabling object recognition in an end-to-end hierarchical fashion. The uptake of CNN-based methods in the domain of remote sensing has been rapid over the past few years, with great potential in numerous applications (Yuan et al., 2020), such as car detection (Ammour et al., 2017), road network delineation (Cheng et al., 2017), remotely sensed scene classification (Zou et al., 2015), land cover and land use classification (Zhang et al., 2018b, 2019b) and semantic segmentation (Kemker et al., 2018).

Semantic segmentation is one of the most fundamental and challenging tasks in remote sensing, where each pixel of the remotely sensed image is assigned a specific thematic label through an automatic process (Marmanis et al., 2018). The goal of semantic segmentation can be either pixel-wise (commonly defined as image classification into e.g. a land cover map) or automated annotation of specific objects of interest. Deep CNNs were designed initially to learn and transfer semantic representations of an image scene into a single value image-wide category (e.g. 'airport', 'residential', 'commercial') through a patch-based training process. This same strategy was adopted for pixel-wise image classification of land cover through application of densely overlapping patches (Fu et al., 2017). However, such a patch-wise procedure introduced smoothing and artefacts at the borders of the classified patches (Zhang et al., 2018b), and the use of overlapping patches involved significant redundant computation. The first deep learning architecture designed for semantic segmentation was the so-called Fully Convolutional Networks (FCN), which represented a breakthrough towards pixel-wise dense labelling of remotely sensed imagery (Volpi and Tuia, 2017). An issue with FCNs (and their extensions) was the trade-off between incorporating a large spatial context and loss of fine spatial detail (e.g. precise boundary delineation) (Marmanis et al., 2018). Therefore, FCNs tended to over-smooth the object, and often involved using a post processing conditional random field (CRF) to sharpen the classification boundaries (Kemker et al., 2018).

A significant contribution to semantic segmentation is the U-Net architecture (Falk et al., 2019) derived from biomedical imaging community. The U-Net introduced an encoder-decoder paradigm, coupled with a set of skip connections within the network, to retain the fine spatial detail after up-sampling procedure (Huang et al., 2018). U-Net has been applied in the field of remote sensing for semantic segmentation to solve different real-world problems, such as extracting road structures (Zhang et al., 2018a, 2018b; Zhang et al., 2018d), surface water bodies (Feng et al., 2019), or mapping natural hazards (Bai et al., 2018). In addition, some researchers also designed new architectures based on U-Net by combining other machine learning and deep learning techniques. For example, Yue et al., (2019) developed a TreeUNet based on a deep U-Net architecture as a set of hierarchical trees for aerial image segmentation. Zhang et al. (2018b), Zhang et al. (2018d) replaced the standard plain U-Net into a Residual U-Net (Res U-Net) through residual networks to cope with the challenges in training and the vanishing (or exploding) gradients (Gao et al., 2019).

Although tremendous progress has been made in deep learning based remotely sensed semantic segmentation, the architecture within deep networks requires an a priori choice of input image patch size (Kemker et al., 2018). However, different landscape objects have their own specific sizes and shapes, and even those objects that are of the same classes may vary distinctively under different socio-ecological conditions (Zhang and Atkinson, 2016). An "optimal" scale is challenging to be determined that is fit-for-purpose across heterogeneous landscapes, and in some cases a single solution may not exist (Graham et al., 2019). An overly large (or small) input scale would negatively affect the classification accuracy for those landscape objects with small (or large) features (Zhang et al., 2018c). To circumvent these scale issues, CNN networks with multiple scales were introduced to characterise feature representations across different spatial scales (Yang et al., 2018). For example, Deng et al., (2018) integrated multiple CNNs with different reception fields to match the scales of objects, thereby achieving increased object classification accuracy. Geng et al. (2020) extracted multi-scale features through pooling and transformation operations, followed by majority voting to increase the accuracy. Zhang et al. (2019a) developed a multi-scale dense network to characterise feature maps at low, middle and high levels, with increased training speed and classification accuracy. Li et al., (2018) transformed features in different scales into the same space, which demonstrated increased accuracy compared with benchmark approaches. Fu et al., (2020) incorporated a rotation-aware and multi-scale CNN for object detection tasks in remotely sensed imagery, which handled effectively the issues of diverse orientation, scales and semantic categories. An object-based CNN that comprised of two distinct scales was also developed to solve the complex task of urban land use classification (Zhang et al., 2018b). Finally, CNNs were utilised to mine the deep features across scales to increase land cover classification accuracy (He et al., 2019). For those multi-scale CNN approaches, a common challenge is the selection of optimal scales (i.e. CNN patch sizes) across a large parameter searching space, where the full range of possible scales is difficult to be explored exhaustively (Sun et al., 2019). In addition, associations amongst the chosen scales were not considered within multi-scale CNNs. To overcome these issues, a scale-sequence joint deep learning method was developed by Zhang et al. (2020) for land use and land cover classifications by integrating continuously increasing scales into the fitting process that are derived autonomously without the need of selecting "optimal" scales manually. However, the joint deep learning framework was designed for classification at different hierarchies (e.g. land cover and land use), which is different from the mapping task at a specific semantic level (e.g. plant species

classification). The knowledge gap is how to incorporate information from multiple scales automatically into the semantic segmentation and classification of individual plants with diverse sizes and shapes.

Here we adopted the scale-sequence principle into the deep residual U-Net as a novel Scale Sequence Residual U-Net (SS Res U-Net) approach and applied it on ultra-fine resolution UAV-based imagery to accurately identify and map individuals of a specific plant species within their natural environment. The SS Res U-Net benefits from information across a sequence of scales, with the ability to characterise the semantic information of each individual plant through precise boundary delineation. The major contribution is to develop a novel, effective multi-scale deep learning approach for remotely sensed segmentation and classification by deriving scales automatically and fitting these into the decision fusion process efficiently. The proposed, novel method was tested on UAV images that are heterogeneous and complex, and benchmarked against different state-of-the-art deep learning methods in terms of both classification accuracy and computational efficiency. The aim is to demonstrate best practice and provide concrete evidence that deep learning can be combined with UAV imagery for mapping and identifying individual plant species.

Our case study uses UAV imagery collected from the Neotropical high-elevation ecosystem known as páramos, and specifically from the páramos of Guantiva - La Rusia, in the departments of Boyacá and Santander, Colombia. Páramos are the world's most biodiverse high-elevation ecosystem (Padilla-González et al., 2017; Cortes et al., 2018) and are found above the timberline in the mountains of Northern South America, usually above 3,200 m. This ecosystem is characterised by the dominance of plants called frailejones, belonging to the subtribe Espeletiinae (family Asteraceae). These plants exhibit a singular growth form, with a stemmed rosette of large, coriaceous leaves, that gives them the appearance of palms, despite belonging to the sunflowers' family. However, their morphology is very diverse, with plant sizes ranging from diminutive fist-sized to more than 15 m tall. Leaves vary from grass-like and smooth to having some of the most densely downy hair found in flowering plants (Diazgranados, 2012; Diazgranados and Barber, 2017). The group is classified in eight genera, with 144 described species, and they are only found in Colombia (89 species), Venezuela (68) and Ecuador (1) (Cuatrecasas, 2013; Diazgranados, 2012; Diazgranados and Barber, 2017).

The páramos of Guantiva - La Rusia have a remarkable richness of frailejones, with 21 reported species, belonging to the genera *Espeletia* (15 species), *Espeletiopsis* (4), *Coespeletia* (1) and *Paramiflos* (1). Nine of the species found here are reported as endangered, including four critically endangered (Diazgranados, 2017; Diazgranados and Castellanos, 2020).

Frailejones are essential for both ecosystems and local inhabitants ecologically and culturally, and are considered keystone species for the conservation of these environments (Cortés et al., 2018; Diazgranados, 2012; Diazgranados and Barber, 2017). These plants are iconic and protected, and are currently under pressure from climate and environmental changes (Varela et al., 2017). A cost effective approach for mapping frailejones across the páramos would enhance the current monitoring of the park authorities and other agencies responsible for the conservation of these plants and their environments.

2. Method

2.1. Study area and data resources

The imagery for our case study was collected from two micro-catchments within the páramos of Guantiva - La Rusia in the departments of Boyacá and Santander, Colombia (Fig. 1). These páramos include a large number of frailejones species (21), with different sizes and leaf induments (from silvery to white or golden trichomes, or even glabrous leaves), giving the rosettes various appearances. Both their diverse morphology and the heterogeneous plant landscape in which they reside make their automatic detection from imagery highly challenging. Here, we focussed on identifying and mapping the dominant species in two study sites, chosen to test the deep learning algorithm: páramo de Ture (S1: vereda Ture, mun. Coromoro, dep. Santander; lat. 6.14360, lon. −72.87767; elev. 3.820–3.840 masl); and páramo de Pan de Azúcar (S2: vereda San Antonio Norte, mun. Duitama, dep. Boyacá; lat. 5.92440, lon. −73.03042; elev. 3,710–3,730 masl). Drone images from the two sites registered four species of frailejones: two acaulirosula species (AR: with no visible stem, short rosettes), i.e. *Espeletia boyacensis* Cuatrec. and *E. congestiflora* Cuatrec.; and two caulirosula species (CR: stemmed or tall rosettes), i.e. *E. incana* Cuatrec. and *E. rositae* Cuatrec. The two AR species are found commonly in the area (primarily *E. congestiflora*), in contrast with the two CR species, which are usually restricted to wetter areas (Table 1).

The S1 site is a relatively simple landscape containing homogeneous matrices of a single class of frailejones species (AR) mixed with grasses, forbs, herbs and mosses. Along the drier slopes, *E. congestiflora* (with a yellowish rosette) dominates with a few sparse individuals of *E. boyacensis* (with a whitish-silvery rosette). The S2 site is a more complex landscape where three frailejones species with two classes (AR

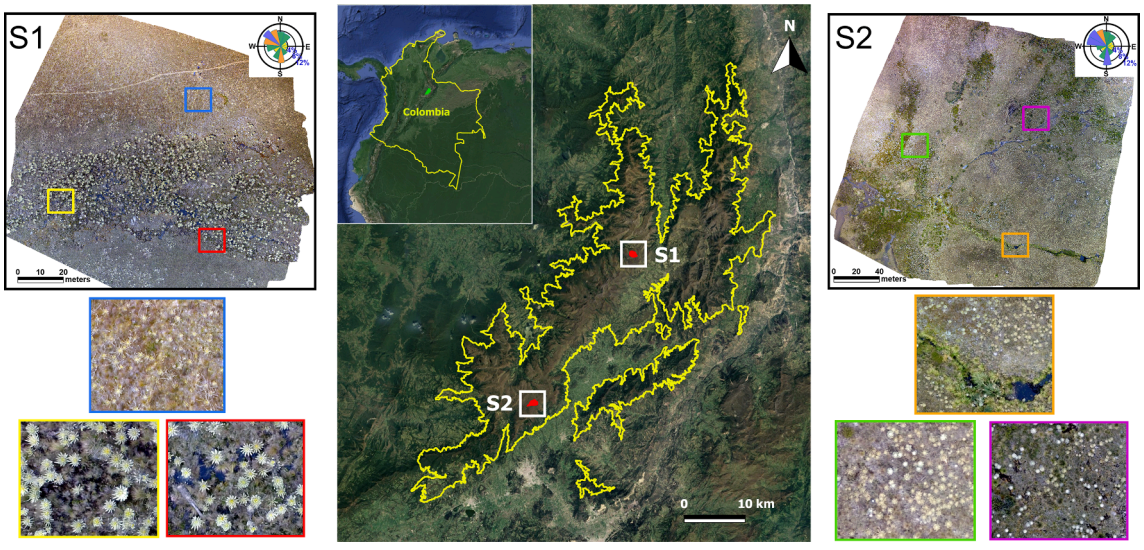

Fig. 1. Two study areas in Colombia: Páramo de Ture (S1) and Páramo de Pan de Azúcar (S2) with typical frailejones species highlighted.

Table 1
The frailejones species found in the study sites with their descriptions and ground photos of typical plants.

Classes	Species	Study sites	Descriptions	Typical ground photos
AR	*Espeletia boyacensis* Cuatrec.	S1, S2	Whitish-silvery rosette plants, 0.7–1.0 m tall, with very short or inconspicuous stem, and large, spreading, yellowish or golden inflorescences with many heads (up to 120). Common names: frailejón, frailejón plateado, frailejón plateado boyacense.	
CR	*Espeletia congestiflora* Cuatrec.	S1, S2	Yellowish rosette plants, 0.7–1.2 m tall, with inconspicuous stem, and very large and robust yellowish or golden inflorescences, with 3–7 congested yellow heads. Common names: frailejón, frailejón de bastón.	
	Espeletia incana Cuatrec.	S2	Whitish rosette plants, up to 8 m tall, with short inflorescences of 3 heads of yellow flowers. Common names: frailejón, frailejón blanco.	
	Espeletia rositae Cuatrec.	S2	Cinereous-yellowish rosette plants, up to 5 m tall, largely surpassed by golden inflorescences of a few (1–5) large dropping yellow heads. Common names: frailejón, frailejón de Santa Rosita.	

and CR) are mixed jointly within a heterogeneous distribution of shrubs, sedges, grasses, forbs, herbs, mosses and bare rock, and where the shrubs and sedges favour the wetter parts. The same two AR species from S1, form matrices in S2 as well, but CR species (*E. incana* and *E. rositae* (with a cinereous-yellowish rosette)) were also found in S2. The heterogeneous and complex landscape of site S2 makes it very well suited for investigating the general applicability of the method.

Aerial photography was acquired by a DJI Phantom 3 UAV drone mounted with a built-in RGB camera. Continuous flights, set to capture a 60% overlap between adjacent image frames, were undertaken on 18 February 2019 and 20 February 2019, providing imagery for sites S1 and S2, respectively. For each site, the image frames were mosaicked, orthorectified and georeferenced to the WGS 1984 projection using UTM Zone 18 N. The dimensions of the resulting mosaicked images were determined by the flight altitude and the speed of the UAV, and were 17,401 × 16,874 pixels with an average spatial resolution of 1.42 cm (i.e. pixel size) for S1 and 12,893 × 11,896 pixels with an average spatial resolution of 2.84 cm for S2. A total of 1140 and 880 plant reference sample points were collected for S1 and S2, respectively, through photo interpretation. Sample points for training, validation and testing were selected using a stratified random scheme, and split into 50% training (S1: 570; S2: 440), 10% validation (S1: 114; S2: 88), and 40% testing (S1: 456; S2: 352). The training samples were used to train the model parameters (weights) and the validation samples to control the design of the network architecture (especially the number of layers within deep networks) and hyper-parameter selection to reduce model over-fitting (Zhang et al., 2019b; Zhang et al., 2019c). The testing samples were used to assess the accuracy. These chosen samples were further checked by cross-referencing to other data sources, such as the Global Biodiversity Information Facility (GBIF), Google Maps, Microsoft Bing Maps, to ensure their precision and validity. Note that most individuals of *E. congestiflora* were in blossom (see the clusters of flowers at the end of the spiderlike tendrils coming out of the rosettes in Fig. 1, blue square), and individuals of frailejones showed young leaf growth in the centre of their rosettes (see the yellow centre in the rosettes in Fig. 1, yellow and red square), all useful features for their automated identification.

2.2. U-Net

A U-Net is an end-to-end deep network composed of an encoder and a decoder that formulates a "U" shape architecture, which has been used for biomedical image segmentation (Ronneberger et al., 2015). The encoder part is a standard convolutional neural network (CNN) with convolutional layers applied repetitively, each followed by a rectified linear unit (ReLU) activation and a max-pooling operation to extract highly generalised feature maps. The decoder part consists of deconvolutions through up-sampling to recover the original resolution and ReLU activations. In addition, the architecture involves several skip connections with the feature maps bypassing the bridge layer with compressed bottleneck embedding, such that the data are not only recovered from a compressed latent representation through up-sampling, but also concatenated directly from the encoder feature map at the same spatial resolution to compensate low-level fine details to high-level semantic feature representations.

2.3. Residual blocks

Residual neural networks proposed by He et al. (2016) involve a stack of residual blocks with residual function and identity mapping through skip connections. Each residual block can be formulated as

$$x_{l+1} = I(x_l) + F(x_l, w_l) \quad (1)$$

Where x_l and x_{l+1} are input and output vectors of the lth residual block, respectively. $I(\cdot)$ is an identity mapping function to copy the input feature maps without learning processes (i.e. $I(x_l) = x_l$). The residual function $F(\cdot)$ is learned from the parameter of w_l of the residual block through backpropagation. He et al., (2016) introduced multiple combinations of batch normalisations, ReLU activations and convolutions within the residual function to facilitate training towards extremely deep networks, such as ResNet-101 with 101 layers. In our case, the computational complexity and actual performance are balanced to simplify the structure by alternating two convolutions and ReLU activations without using the batch normalisation layers.

2.4. Residual U-Net (Res U-Net)

Residual U-Net (Res U-Net) is established by replacing the plain convolutions and ReLU activations within the standard U-Net into the residual blocks. In this way, the advantages of U-Net and residual networks are integrated together for semantic segmentation (Zhang et al.,

2018d). The key benefit of Res U-Net is to focus the training process on the residual instead of the whole network, thus addressing network degradation issues effectively (Diakogiannis et al., 2020). It also incorporates skip connections within and between residual blocks, where information is allowed to propagate from low-level features to high-level semantics. Similar to the standard U-Net, the Res U-Net also involves decoder and encoder parts (Fig. 2). The decoder part has two residual blocks with two 3 × 3 convolutional layers and ReLU activations followed by an identity mapping. Instead of using a 2 × 2 max-pooling layer like traditional CNNs, the first convolutional layer here has a stride of 2 to sub-sample the feature maps along with the convolutional process. Likewise, the encoder part also has two residual blocks, and the connections between the encoder part as well as between the residual blocks use the up-sampling layers to recover the fine spatial resolution (Fig. 2). Note, the unnecessary cropping step in the standard U-Net is removed in the network. Between the decoder and encoder parts, there is also a bridge using a residual block to obtain the compressed bottleneck representations. The final output layer is a softmax regression layer to achieve the final classification results.

2.5. Scale sequence Residual U-Net (SS Res U-Net)

SS Res U-Net utilises Residual U-Net (Res U-Net) to classify a particular class or multiple classes of an image M through iteration, with input scales in sequence (from small to large) defined by a scale sequence denoted as $\theta = \{\theta_1, \theta_2, \ldots, \theta_i, \ldots\theta_n\}$ (Fig. 2). Here, θ_1 and θ_n refer to the minimum and maximum scale, respectively, which are computed based on the minimum and maximum geometric sizes of the objects within a class or classes; n is the total number of scales; θ_i ($i = 1,2, \ldots, n$) refers to the i-th scale that is derived from linear interpolation between minimum and the maximum scales by specifying the total number of scales (Zhang et al., 2020). Thus i and n also refer to the order of sequence and the total number of iterations required for model predictions. The specific task of classifying an individual of frailejones (denoted as class *Frai*) is used as an example to develop the proposed SS Res U-Net hereafter.

The process of classifying individual frailejones is an iterative process. For $i = 1$ (i.e., the first iteration) the membership degree (probability) of being a frailejones in the image M can be predicted using the Res U-Net as follows:

$$P(Frai^1) = Res\ U\ -\ Net.\ Predict(\theta_1, M) \tag{2}$$

The first classification probability $P(Frai^1)$ of frailejones at the first scale θ_1 in the scale sequence can be derived from the probabilistic prediction of Residual U-Net (Res U-Net). For the ith scale, the corresponding classification results $P(Frai^i)$, will be integrated with the spectral value of the remotely sensed image (M) to achieve the integrated data of the ith scale (denoted as $FraiData^i$), which will be used as the input data for the next iteration as follows:

$$FraiData^i = Concate(M,\ P(Frai^i)) \tag{3}$$

Where, the function *Concate* (*) in Eq. (3) refers to the integration of image M and $P(Frai^1)$ at the band dimension.

In the following iterative process ($i = 2, \ldots, n$), the probability of frailejones is formulated as a Markov Chain Process, where the predictions of the i-th iteration are affected by the output classification results of the previous iteration. Meanwhile, the membership values (i.e. classification probabilities) of frailejones are acquired in each iteration through:

$$\begin{aligned} P(Frai^i) &= P(Frai(\theta_i)^i | P(FraiData^{i-1}) \\ &= Res\ U - Net.Predict(\theta_i, FraiData^{i-1}) \end{aligned} \tag{4}$$

Eq. (4) shows that the membership probabilities of frailejones predicted by Res U-Net model at ith scale, which are affected by the

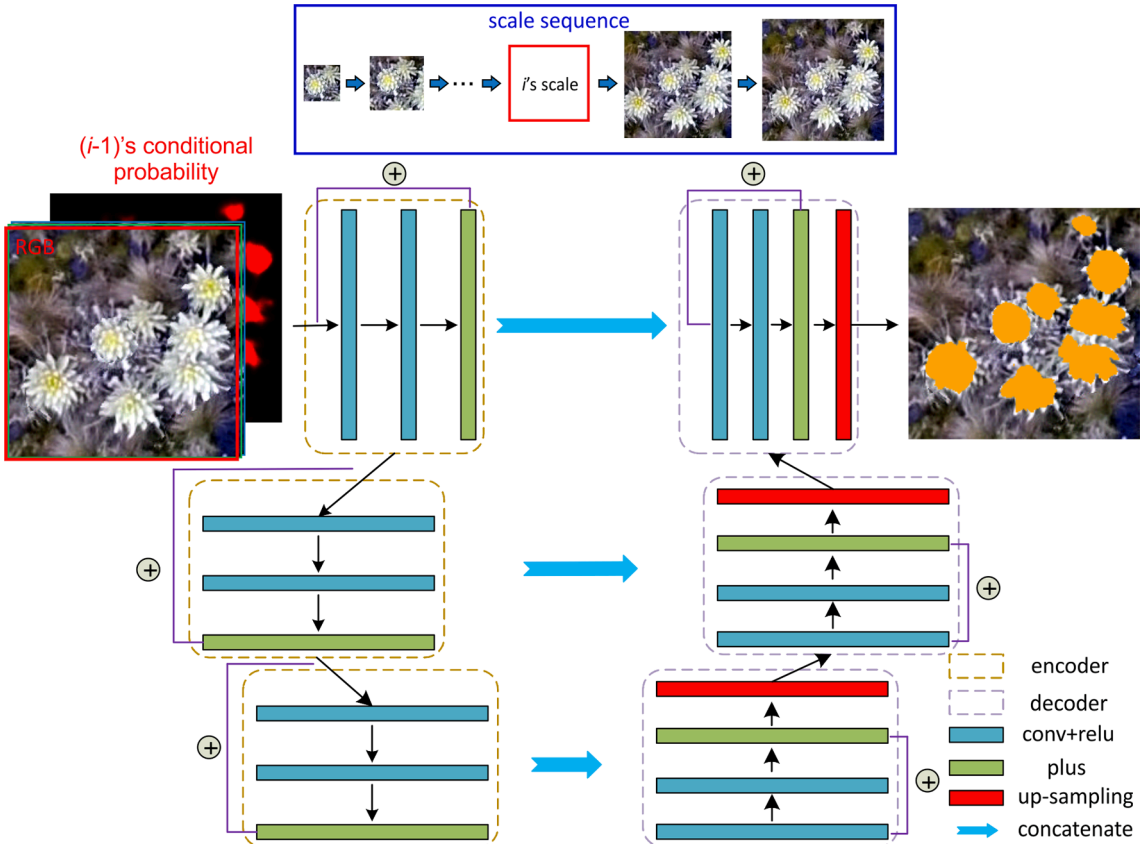

Fig. 2. Model architecture of the Scale Sequence Residual U-Net.

$FraiData^{i-1}$ acquired from the previous iteration using Eq. (3). The final classification map ($M_{\text{Frairesult}}$) is achieved by taking the prediction of the maximum probability of the final classification probability $P(Frai^n)$:

$$M_{\text{Frairesult}} = argmax(P(Frai^n)) \quad (5)$$

2.6. Quantitative evaluation method on the SS Res U-Net

Benchmark classifiers: In addition to the Standard U-Net and Res U-Net as mentioned above, two other common deep learning-based classification methods (fully convolutional networks (FCN) and Multi-Scale Res U-Net (MS Res U-Net)) were used as benchmarks to test the performance of the proposed SS Res U-net in this research. Both FCN and MS Res U-Net are introduced briefly hereafter.

FCN: A Fully Convolutional Network (FCN) is one of the most basic frameworks for pixel-wise semantic segmentation (Long et al., 2015). It relies on the classical patch-based CNN to extract high-level feature representations through convolutional and pooling layers as an encoder, and replaces the fully connected layers (used for predicting a single label at patch level) into a set of deconvolution operations as a decoder (used for up-sampling the prediction onto the original spatial resolution) (Volpi and Tuia, 2017). The major difference between FCN and the standard U-Net is that the FCN does not involve skip connections between the corresponding encoder and decoder layers.

MS Residual U-Net (MS Res U-Net) utilises multiple Residual U-Nets across different scales as a combination, whose basic model structure and architecture is exactly the same as the single-scale Residual U-Net. In this research, three scales were adopted to capture the scale effects by varying the CNN input window sizes, and a majority voting strategy (Lv et al., 2018) was employed to achieve final classification results.

To make a fair comparison, the corresponding model structure and parameters of all these benchmarks were maintained to be as consistent as possible with the proposed SS Residual U-Net.

Metrics for quantitative evaluation: The classification results were assessed by a confusion matrix, comprised of the number of pixels allocated as true positive (*TP*), true negative (*TN*), false positive (*FP*), and false negative (*FN*). Thematic accuracy metrics were used to evaluate quantitatively the classification performance, including overall accuracy (*OA*), precision (*P*), recall (*R*), F1 score (*F*1) (Bayr and Puschmann, 2019). These metrics are derived as follows:

$$P = TP/(TP + FP) \quad (6)$$

$$R = TP/(TP + FN) \quad (7)$$

$$F1 = 2 \times (P \times R)/(P + R) \quad (8)$$

$$OA = (TP + TN)/(TP + FN + TN + FR) \quad (9)$$

In addition, the geometric accuracy was also tested through an Intersection over Union (*IoU*) metric, which measures how well the segmentation matches the corresponding ground reference (Falk et al., 2019). Specifically, the IoU was calculated by dividing the number of pixels labelled as a segment/object in both prediction and reference (intersection of two segments), by the number of pixels labelled as that segment/object in either the prediction or in the reference (union of two segments).

3. Experimental results and analysis

3.1. SS Res U-Net model parameter settings

Model parameters for the proposed SS Res U-Net involved setting the minimum scale and maximum scale, as well as the number of iterations as a sequence of scales throughout the process. Details are demonstrated hereafter.

3.1.1. Minimum and maximum scales

Both minimum and maximum scales are determined based on the smallest and largest size of image objects (i.e. individual frailejones) present, obtained by object-based image segmentation. Given the difference in spatial resolution between S1 (1.42 cm) and S2 (2.84 cm), the minimum and the maximum scales of the scale sequence in S1 are different from those in S2. In S1, the minimum and maximum input window sizes of the scale sequence were set to 16 × 16 and 144 × 144 respectively, based on the size of the smallest (about 22 cm) and largest object (about 198 cm) captured in the image. The maximum scale is purposefully slightly larger than the size of the largest object. This is to capture the spatial context of the species within the scene. Similarly, the minimum and maximum window sizes of the scale sequence in S2 were set to 12 × 12 and 108 × 108, to take into account the smallest (32 cm) and largest (307 cm) of the segmented objects in the image.

3.1.2. Number of iterations in scale sequence Residual U-Net

A range of scales are linearly interpolated into the network to achieve a scale sequence between minimum and maximum scale using different input CNN window sizes. The smallest iteration possible is two, which only includes the minimum and maximum scales, and the number of iterations increases proportional to the chosen number of scales. Using the validation samples, we optimised the number of iterations and sequence of scales by maximising the mapping accuracy. Fig. 3 shows the progression of the overall accuracy for small subareas of both S1 and S2 with increasing number of iterations. For both study sites, the accuracy initially increases rapidly from around 87%, with two iterations only, towards approximately 91.5% after five iterations. The accuracy then increases further slowly and asymptotically with increasing iteration. The five iterations involve five input CNN window sizes as a sequence of scales, including 16 × 16, 48 × 48, 80 × 80, 112 × 112, and 144 × 144 for S1, and 12 × 12, 36 × 36, 60 × 60, 84 × 84, 108 × 108 for S2, respectively. Fig. 4 shows the intermediate classification results of S1 and S2 along the first five iterations (2 – 6). The classification process of frailejones starts from the centre of the plant rosette at the smallest scale with CNN window sizes of 16 and 12 for S1 and S2 respectively (Fig. 4b) and as the input window size increases across the scale sequence, it gradually grows to the edges of the plant rosette (Fig. 4c to 4f), nearly exactly matching the reference (Fig. 4a).

3.1.3. Accuracy comparison across different CNN window sizes

The proposed SS Res U-Net involves fitting using a sequence of scales (i.e., image patch sizes) throughout the process without optimal scale selection. All four benchmarks, including the FCN, standard U-Net, Res U-Net, and the MS Res U-Net involve significant effort in terms of scale selection across a wide range of CNN window sizes (16 × 16, 32 × 32, 48 × 48, 64 × 64, 80 × 80, 96 × 96, 112 × 112, 128 × 128, and 144 × 144) in S1 and (12 × 12, 24 × 24, 36 × 36, 48 × 48, 60 × 60, 72 × 72, 84 × 84, 96 × 96, 108 × 108) in S2. MS Res U-Net requires a further step to determine the best combination of three window sizes (scales). In this research, the three input CNN window sizes of MS Res U-Net were chosen across those candidate window sizes through trial and error. The best resulting combination for S1 was 48 × 48, 96 × 96 and 112 × 112 and for S2 was 24 × 24, 48 × 48 and 84 × 84. To further demonstrate the scale selection process, a set of scales (CNN input window sizes) were tested with 20 iterations for each scale as benchmark comparison against the proposed SS Res U-Net with five iterations only (Fig. 5). Clearly, the SS Res U-Net achieved the greatest accuracies (91.89% and 89.65%) (solid blue lines), higher than MS Res U-Net (88.35% and 87.42%) (dash red lines) and all the other three benchmarks (Fig. 5).

3.2. Classification results and analysis

The SS Res U-Net approach was implemented with the optimal number of iterations (five) and scale range, and classifications were compared with those achieved by the benchmark methods (FCN,

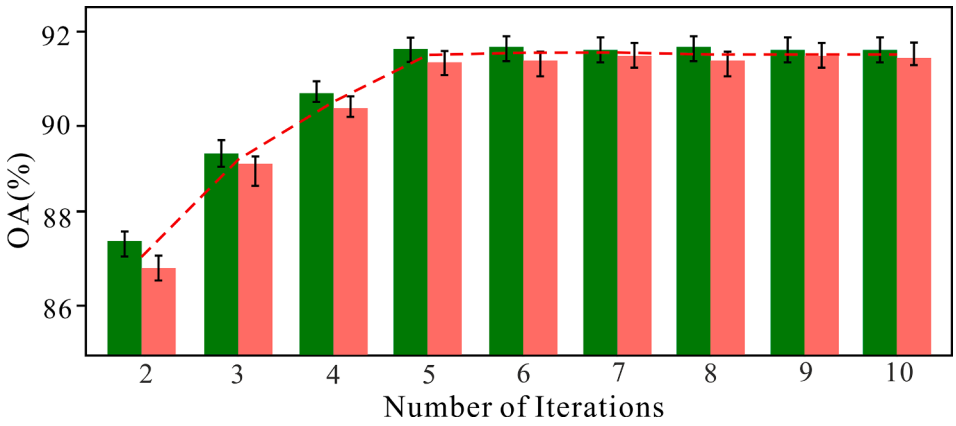

Fig. 3. The progression of the overall accuracy (OA) for a subarea of S1 (in green) and S2 (in red) with increasing number of iterations (2–10). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

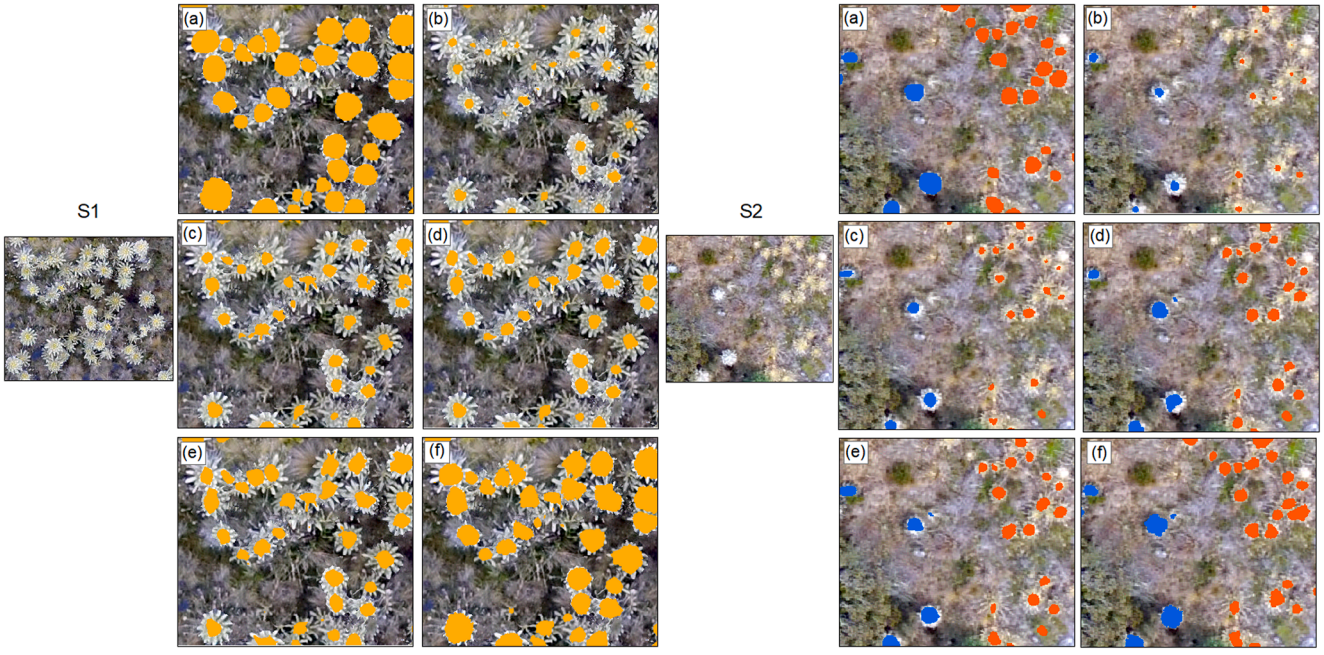

Fig. 4. Illustration of the classification results with increasing iterations for a subarea of S1 (left) and S2 (right): (a) the ground reference for validation; and the classification results for (b) 2 iterations, (c) 3 iterations, (d) 4 iterations, (e) 5 iterations, and (f) 6 iterations.

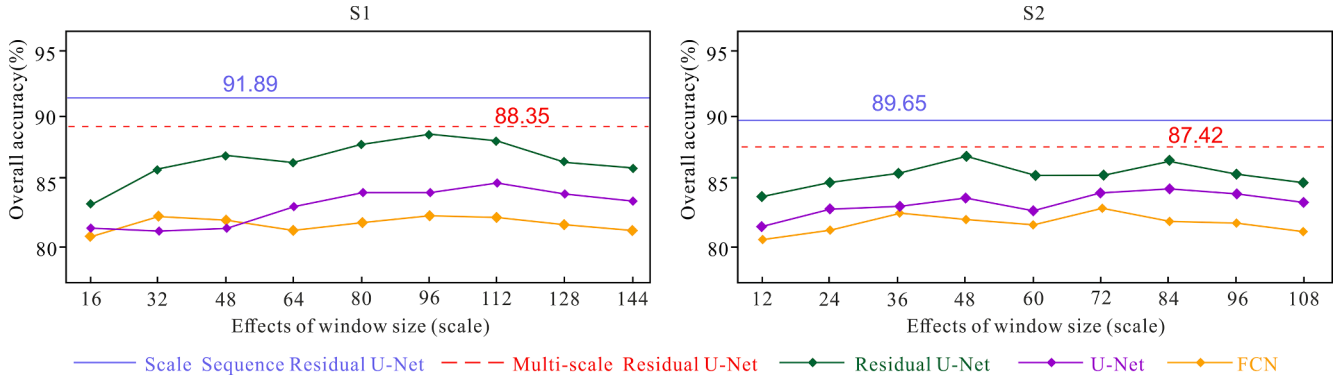

Fig. 5. The effects of CNN input window sizes (scales) on the classification accuracies in S1 and S2 for the proposed Scale Sequence Residual U-Net and the four benchmark approaches.

standard U-Net, Res U-Net and MS Res U-Net). The classification results were evaluated through visual inspection and a quantitative assessment, involving semantic accuracy (overall accuracy (OA), precision, recall, F1) and geometric accuracy (IoU).

3.2.1. Classification results, visual inspection

The classification results of the SS Res U-Net and the four benchmark comparators are shown for 4 image subsets of S1 (Fig. 6). It can be seen that SS Res U-Net differentiated each individual frailejones successfully with accurate boundaries between plants (in yellow circles), while the benchmark methods universally failed to separate these individual plants with substantial mistakes (in red circles). Fig. 6(a) and 6(b) show the classification results for densely distributed frailejones. Clearly, both FCN and U-Net merged individual frailejones into some continuously distributed patches, such as in the upper right of Fig. 6(a) and upper left of Fig. 6(b). Res U-Net achieved a higher accuracy by correctly characterising the edges of frailejones, but failed to separate individuals where plants were densely packed (see red circles in the lower right side of Fig. 6(a) and the bottom of Fig. 6(b)). MS Res U-Net showed an improvement in both detecting and characterising individual plants, but could still not effectively handle densely packed plants (e.g. the red circle in the upper left of Fig. 6(b)). SS Res U-Net demonstrated a significant improvement in capturing individual frailejones with precise boundary delineation and segmentation as highlighted by yellow circles in Fig. 6(a) and 6(b). Where Frailejones are sparser and within a dense (Fig. 6(c)) or sparse (Fig. 6(d)) cover of other types of vegetation, both FCN and U-Net tended to incorrectly classify shrubs (Fig. 6(c)) or grass tufts (Fig. 6(d)) into frailejones and had issues with separating frailejones which are close to each other (red circles in Fig. 6(c) and 6(d)). Res U-Net and MS Res U-Net delivered similar and better classification results, while still struggled to separate neighbouring frailejones (red circles in Fig. 6(c) and 6(d)). In contrast, SS Residual U-Net identified the frailejones accurately, with individual plants differentiated clearly (yellow circles in Fig. 6(c) and 6(d)).

Fig. 7 shows the classification results for four subsets in S2. Fig. 7 (a – c) show results for areas containing both densely distributed AR species and sparsely distributed CR species within a matrix of shrubs, water and grass. Fig. 7(d) shows an area containing CR species within a matrix of shrubs, water and grass. As illustrated by the Fig. 7(a - d), the SS Res U-Net approach is more superior than the benchmarks in identifying both tall and short species of frailejones (AR in orange and CR in blue) and separating individual plants.

The approaches' performances varied slightly with frailejones species: Individual AR plants were often merged into continuous patches by the FCN (top right red circle in Fig. 7(a) and bottom red circle in Fig. 7 (c)), and the geometric shapes of the classified plants were seriously distorted (e.g. the right middle red circle in Fig. 7(b)). Similarly, the classification results of the standard U-Net showed severe geometric distortions, merging the AR plants together when they are close to each other (see upper right red circles in Fig. 7(a) and 7(b) and the bottom red circle in Fig. 7(c)). Res U-Net revealed some benefits in characterising the geometry of the AR species through segmentation, but still struggled to separate dense, continuously distributed plants (e.g. upper right circle in Fig. 7(a) and bottom red circle in Fig. 7(c)). MS Res U-Net showed a significant improvement in the detection and geometric characterisation of both species, with some accurate results in classifying AR species. Nevertheless, it produced poorly defined boundaries in areas of high Espeletia plant density (e.g. bottom red circles in Fig. 7(c)). SS Res U-Net achieved better results across the board showing an increased accuracy in differentiating between adjacent AR plants and in capturing their boundaries precisely.

For CR species, FCN and U-Net performed similarly, producing severely merged patches (red circles in Fig. 7(a), 7(b) and 7(d)) and in some cases omitting small sized plants (e.g. upper right red circle in Fig. 7(c)). Res U-Net showed some improvement in characterising the geometry of the CR species, but in some cases would still merge nearby plants into a single patch (e.g. middle red circle in Fig. 7(a), upper left red circle in Fig. 7(b) and red circles in 7(d)). MS Res U-Net showed an improvement in separating individual plants in some cases (e.g. yellow circles in Fig. 7(b) and 7(d)), while in other cases would still merge densely distributed plants into large patches (e.g. middle red circle in Fig. 7(a) and red circle in 7(d)). SS Res U-Net solved all the above problems and classified the CR plants individually and accurately through precise detection and segmentation (yellow circles in Fig. 7(a-

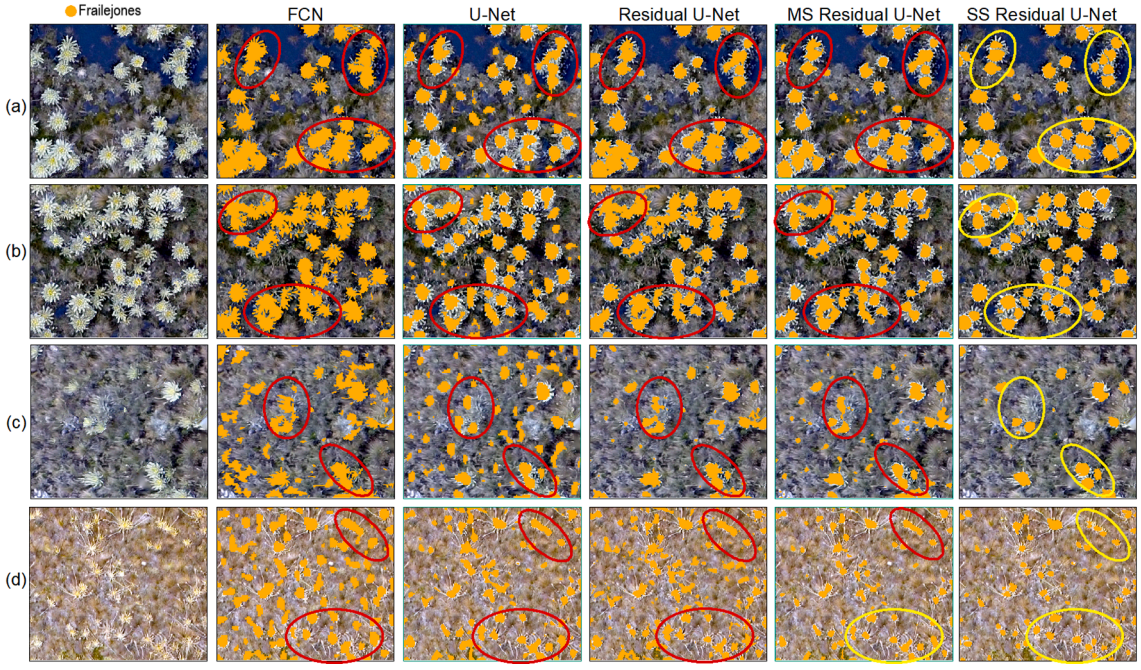

Fig. 6. Four subsets (i.e., a, b, c and d) of frailejones classification in S1 using FCN, U-Net, Residual U-Net, MS Residual U-Net, and the proposed SS Residual U-Net. The yellow and red circles represent correct and incorrect classifications results, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

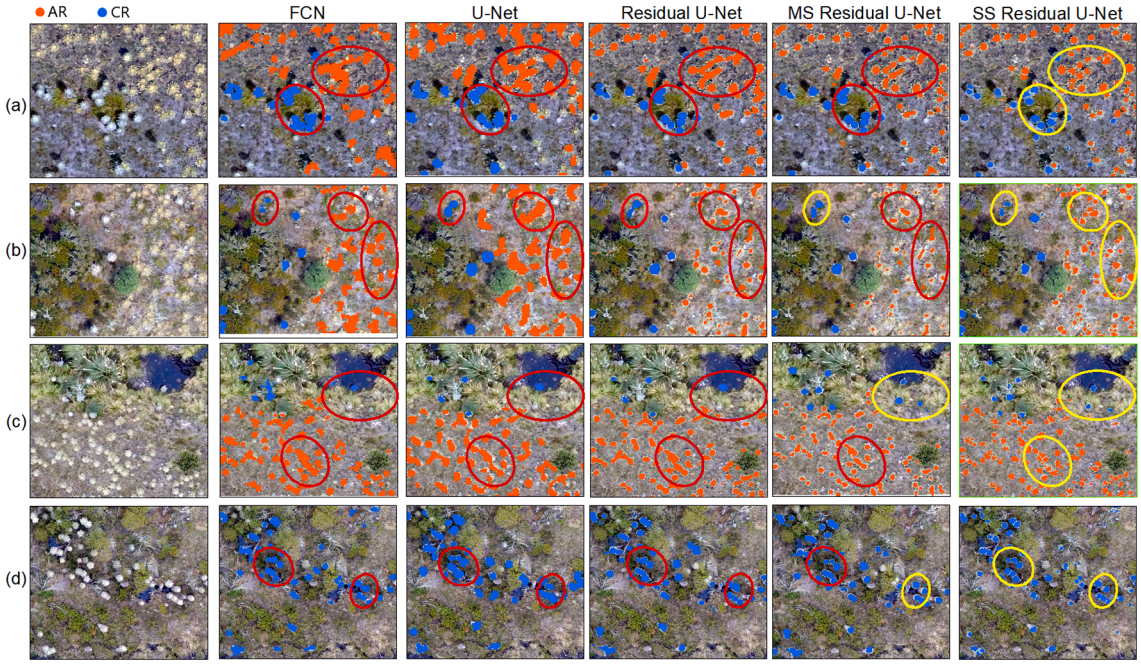

Fig. 7. Four subsets (i.e., a, b, c and d) of frailejones (AR: orange; CR: blue) classification in S2 using FCN, U-Net, Residual U-Net, MS Residual U-Net, and the proposed SS Residual U-Net. The yellow and red circles represent correct and incorrect classifications results, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

d)). It also successfully captured small plants (top yellow circle in Fig. 7 (c)).

3.2.2. Classification result, quantitative accuracy assessment

The effectiveness of the SS Res U-Net was further demonstrated by a quantitative assessment of the classification accuracies. For both study sites (Tables 2 and 3), the proposed SS Res U-Net achieved the largest overall accuracy for the classification of individual frailejones in S1 (i.e. 91.89%) and of individual frailejones and species in S2 (i.e. 91.45%). This is significantly higher than the accuracies achieved by the four benchmarks. The proposed SS Res U-Net also achieved the highest precision, recall and F1 scores for both study areas and the different frailejones species. Moreover, the geometric accuracy of the intersection over union (IoU) was also the greatest for the SS Res U-Net (S1: 89.64% and S2: 90.34%) and overall accuracy, detection rate and geometric accuracy results between study sites were very similar.

Amongst the benchmarks, the best performing approach was MS Res U-Net, followed in sequence by Res U-Net, the standard U-Net, and the FCN.

3.2.3. Influence of training sample size on classification accuracy

To test the sensitivity of the proposed SS Res U-Net method and the four benchmarks approaches to training sample size, the approaches were run using randomly selected training subsets representing 90%, 70%, and 50% of the original training sample. Fig. 8 shows the reduction in classification accuracy of the approaches with reducing training sample size. SS Res U-Net is the least affected by sample size for both study sites S1 and S2, showing the smallest reduction in accuracies: about 2.5%, 12% and 19% when using 90%, 70% and 50% of the original sample size respectively. Res U-Net and MS Res U-Net behaved similarly and showed slightly higher sensitivity than SS Res U-Net, with decreases in accuracy of about 4%, 14.5% and 23% for the 90%, 70% and 50% samples respectively. The U-Net revealed a further decrease in accuracy, but mainly for the 50% sample (reductions of about 4.6%, 15.5% and 33%, respectively). The most sensitive approach was FCN, where largest accuracy decreases were observed of about 6%, 18% and 37% respectively.

Table 2
Accuracy assessment in S1 for the five methods (FCN, Standard U-Net, Res U-Net, MS Res U-Net and SS Res U-Net) using overall accuracy (OA), precision, recall, F1 score and Intersection over Union (IoU). The bold font represents the highest accuracy for each of the quantitative metrics.

Method	OA	Precision	Recall	F1	IoU
FCN	80.53	78.46	76.39	77.55	82.06
Standard U-Net	85.42	83.28	84.53	83.90	85.32
Res U-Net	86.28	85.39	86.29	84.82	86.15
MS Res U-Net	88.35	86.74	88.08	87.41	87.56
SS Res U-Net	**91.89**	**90.75**	**91.92**	**91.33**	**89.64**

Table 3
Accuracy assessment in S2 for the five methods (FCN, Standard U-Net, Res U-Net, MS Res U-Net, and SS Res U-Net) for the two frailejones species (AR: *E. conglomerata* and CR: *E. incana*) using the overall accuracy (OA), precision, recall, F1 and Intersection over Union (IoU). The bold font represents the highest accuracy for each of the quantitative metrics.

Method	OA	Espeletia Type	Precision	Recall	Mean F1	Mean IoU
FCN	80.29	AR	77.28	79.44	79.38	81.53
		CR	80.62	82.35		
Standard U-Net	84.33	AR	82.65	80.37	82.76	84.92
		CR	83.46	84.73		
Res U-Net	86.39	AR	84.63	85.28	85.28	86.42
		CR	87.52	87.42		
MS Res U-Net	88.08	AR	87.19	88.85	87.64	88.05
		CR	89.62	89.47		
SS Res U-Net	**91.45**	AR	**89.65**	**90.34**	**90.82**	**90.34**
		CR	**92.76**	**91.93**		

3.2.4. Comparison of computational efficiency

The computational efficiency of our method was tested and compared with those of the benchmark approaches (Table 4). The classification experiments were implemented using Keras backend with Tensorflow under python platform through a workstation deployed by NVIDIA GeForce GTX 1080 Ti with inner memory of 32 GB. The

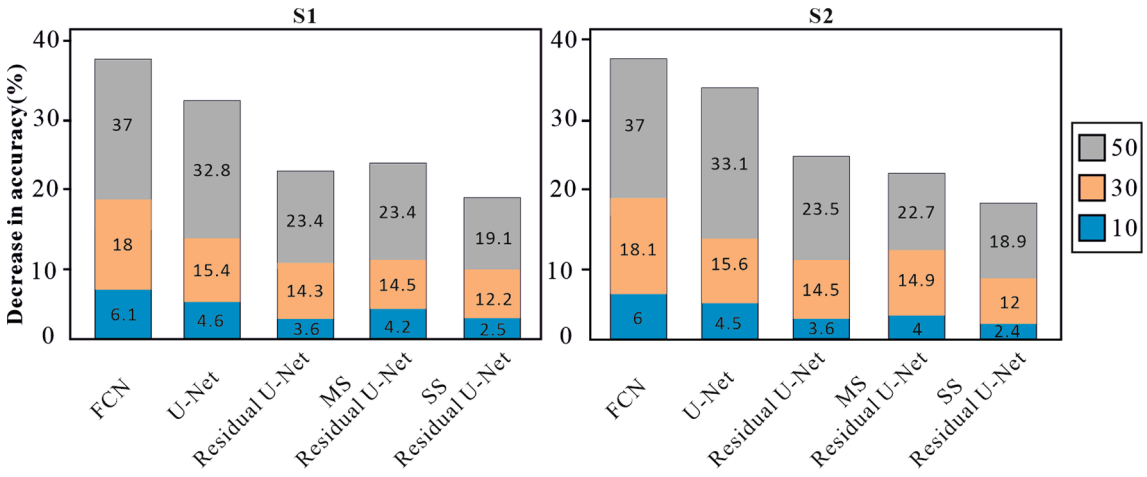

Fig. 8. The effect of sample size (randomly selected subsets representing 90%, 70%, 10% of the original training sample size) on overall accuracy, for the proposed SS Res U-Net method and the benchmark comparators at study sites S1 and S2.

Table 4
Computational time comparison for five methods in S1 and S2. The bold font represents the least running time and the highest computational efficiency.

Method	S1	S2
FCN	12.15 h	12.42 h
Standard U-Net	10.17 h	11.26 h
Res U-Net	8.86 h	9.95 h
MS Res U-Net	18.43 h	20.37 h
SS Res U-Net	**1.28 h**	**1.86 h**

proposed SS Res U-Net involved five iterations, and took only 1.28 h and 1.86 h for S1 and S2, respectively. In contrast, FCN and the other U-Net based benchmarks required significantly more time (as shown in Table 4) to complete the process of scale selection, which involved searching across a wide range of CNN windows to determine the optimal scale and acquire the highest classification accuracy. This process resulted in a remarkable increase in running time, which depended on the complexity of each benchmark and could be up to one order of magnitude in comparison with the proposed SS Res U-Net.

4. Discussion

UAV based remote sensing can produce ultra-fine spatial resolution imagery with highly detailed spatial information content. If fully exploited by image analysis approaches based on deep learning, this could offer a great opportunity to identify, map, count and monitor plant individuals in complex and diverse ecosystems (Masiero et al., 2017). However, despite this potential, there exist huge challenges in handling such abundant information, where the scales of the same ground objects can vary distinctively across different parts of the image (Yang et al., 2018), even within a small imaging region. This is because the flight altitude of a UAV drone is commonly low, with a relatively short distance between the camera and the ground surface. Both the height of ground objects and topographic conditions could have a significant impact on the scale of objects captured in the image. In addition, the data captured from the UAV borne sensors often contain disturbances and noise due to the shaking of the UAV during flying, leading to unwanted variations in UAV imagery (Zhang et al., 2019b; Zhang et al., 2019c). Existing deep learning approaches require substantial effort to determine the "optimal" scale or patch size, which involves the labour-intensive and time-consuming process of scale selection to achieve the highest classification accuracy over a set of training datasets. However, such an "optimal" scale is often a sub-optimal solution, since it is hard to represent in a single scale the variation in size and geometry of image objects across large areas. Multi-scale CNNs can partially represent the scale variation using majority voting from CNNs with three different input window sizes. Yet, the actual patch sizes were either determined empirically (e.g. Sun et al., 2019), or searched exhaustively across a huge parameter space (e.g. Lv et al., 2018). Hence, it has been challenging or even impossible to determine one or multiple optimal scales using current scale selection strategies. In addition to this natural variation, the spatial context of the frailejones species themselves can be highly complex and heterogeneous, particularly for frailejones that are clustered together, and where the edges of leaves and flowers are arranged in "crisscross" patterns. Thus, mapping individual plants, such as the frailejones of our case study, from UAV imagery using existing deep learning-based semantic segmentation methods can produce disappointing results, often resulting in a continuous patch of several adjacent plants together as a whole, without the capability to delineate precise boundaries between them. A novel scale sequence Residual U-Net (SS Res U-Net) was proposed in this research to enhance the ability to learn the boundary of individual plant species through a newly designed strategy to integrate a sequence of scales (patch sizes) over multiple Res U-Net models. The approach was tested to map frailejones which have a complex and varying geometry and are typically found within a diverse and heterogeneous mountain landscape.

The experimental results demonstrated that the SS Res U-Net consistently achieved the highest classification accuracy compared to all benchmark approaches, and showed the least sensitivity to sample size reduction, demonstrating outstanding performance in both classification accuracy and robustness. The proposed SS Res U-Net was superior to not only the Res U-Net (Gao et al., 2019) and U-Net (Zhang et al., 2018d), but also to the multi-scale Residual U-Net (MS Res U-Net). Essentially, the proposed SS Res U-Net is a multi-scale classifier retaining the advantages of the Res U-Net, but performing in a way that sequentially integrates different scales of networks through iteration, with features learnt from a small scale gradually passing into larger scales. Interactions amongst different scales of networks were linked with information transmission along the adjacent scales of representations as a Markov Chain process. Deep features across different scales could then be mined through the re-enforcement learning process, as shown in the experimental results, where the spatial patterns of complex landscapes could be differentiated with accuracy. By contrast, the different scales of CNN models in MS Res U-Net are separated from each other, without capturing their associations or conditional dependencies in the decision fusion process. This should, at least partially, explain the higher accuracy achieved by SS Res U-Net compared with the multi-scale approach (MS Res U-Net). This major issue exists commonly in other state-of-the-art multi-scale CNN approaches (e.g. Fu et al., 2020;

Geng et al., 2020; He et al., 2019; Zhang et al., 2019a). Another benefit of employing the scale sequence strategy is to automate the determination of input scales and to avoid the time-consuming and tedious process of optimal scale selection. For a particular scale sequence, only two parameters involving the minimum and the maximum scales are predefined, which can be derived by the minimum and maximum sizes of objects. As the sizes of image objects were measured directly from the remotely sensed imagery, the corresponding scale sequence was self-adaptive to a huge variation of scales over different images. Real-world features tend to manifest across a range of scales, from small to large, and present with different sizes and geometries. With five iterations only, the maximum classification accuracies were achieved in both study sites, which was in line with results for scale-sequence joint deep learning applied to land use and land cover classification (Zhang et al., 2020). Meanwhile, the classification process of the proposed method was greatly simplified in comparison with the benchmark approaches, leading to a substantially reduced computational time, from over 10 h to just above one hour in the experiments, which can be regarded as generally acceptable to end-users. Simple, fast and accurate classification is the ultimate goal of deep learning research towards operational deployment in the remote sensing community (Gupta et al., 2020).

Although the proposed method was used to classify frailejones species at the individual plant level using UAV images as an example in this research, the method is also relevant to a wide range of other image classification and semantic segmentation tasks using different sources of remotely sensed images (e.g. SAR, optical satellite sensor imagery etc.). The SS Res U-Net employed multi-scale deep learning approaches, without the process of optimal scale selection. This indicates that the proposed algorithm is more consistent with human visual cognition. Human beings have the ability of multi-scale observation, but never spend time selecting optimal scales while observing a real scene or an image (Zhang et al., 2020). Thus, the proposed method coincides with the research objective of machine intelligence, to mimic the human visual system to identify and characterise complex patterns. From an artificial intelligence perspective, the proposed method has a wide application prospect across diverse domains including computer vision and pattern recognition to enhance the representation ability across different scales through scale sequencing.

5. Conclusion

UAV drone imagery has great potential to characterise individual frailejones species at ultra-fine centimetre spatial scale. Yet, the semantic segmentation of such drone images using existing deep learning methods remains challenging, due to variation in the scales, geometries and densities of the objects of interest presented in the images, and the complex landscape patterns captured at fine spatial details. This research proposed a novel Scale Sequence Residual U-Net (SS Res U-Net) method to address this challenge in semantic segmentation, with a specific exemplar of individual frailejones mapping. The SS Res U-Net inherited the high classification accuracy from Residual U-Net at each scale, and further enhanced this through a scale sequence to allow information transmission across continuously increasing scales for precise segmentation and boundary delineation. The input scales within the scale sequence were derived automatically, and different Residual U-Nets were integrated through iteration as a Markov chain process. In this manner, the features learnt at the lowest-level were passed gradually towards high-level representations while maintaining precise boundary characteristics. Experiments using UAV drone images demonstrated that the proposed method consistently outperformed four deep learning benchmarks in terms of adaptation to scale variation, with the highest classification accuracy and robustness, as well as computational efficiency. Especially, the computational time of the SS Res U-Net was significantly reduced to a user acceptable range, thereby greatly enhancing the practical utility of this deep learning technique. In conclusion, the proposed SS Res U-Net is a promising method that is fast, parsimonious, scale-adaptive and highly accurate, and shows strong robustness in classification of remotely sensed imagery over complex landscapes with a wide range of applications in remote sensing and beyond.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research was sponsored by Centre of Excellence in Environmental Data Science (CEEDS), joint venture between UK Centre for Ecology & Hydrology and Lancaster University. The UAV data capture in Colombia was supported by NERC/AHRC funded Newton project PARAGUAS (NE/R017654/1). The authors are grateful to the two anonymous referees for their constructive comments and suggestions on this manuscript.

References

Aasen, H., Honkavaara, E., Lucieer, A., Zarco-Tejada, P.J., 2018. Quantitative remote sensing at ultra-high resolution with UAV spectroscopy: A review of sensor technology, measurement procedures, and data correctionworkflows. Remote Sens. 10 https://doi.org/10.3390/rs10071091.

Ammour, N., Alhichri, H., Bazi, Y., Benjdira, B., Alajlan, N., Zuair, M., 2017. Deep learning approach for car detection in UAV imagery. Remote Sens. 9 https://doi.org/10.3390/rs9040312.

Baena, S., Moat, J., Whaley, O., Boyd, D.S., 2017. Identifying species from the air: UAVs and the very high resolution challenge for plant conservation. PLoS One 12. https://doi.org/10.1371/journal.pone.0188714.

Bai, Y., Mas, E., Koshimura, S., 2018. Towards operational satellite-based damage-mapping using U-net convolutional network: A case study of 2011 Tohoku Earthquake-Tsunami. Remote Sens. 10 https://doi.org/10.3390/rs10101626.

Bayr, U., Puschmann, O., 2019. Automatic detection of woody vegetation in repeat landscape photographs using a convolutional neural network. Ecol. Inform. 50, 220–233. https://doi.org/10.1016/j.ecoinf.2019.01.012.

Blaschke, T., Hay, G.J., Kelly, M., Lang, S., Hofmann, P., Addink, E., Queiroz Feitosa, R., van der Meer, F., van der Werff, H., van Coillie, F., Tiede, D., 2014. Geographic object-based image analysis - towards a new paradigm. ISPRS J. Photogramm. Remote Sens. 87, 180–191. https://doi.org/10.1016/j.isprsjprs.2013.09.014.

Cheng, G., Wang, Y., Xu, S., Wang, H., Xiang, S., Pan, C., 2017. Automatic road detection and centerline extraction via cascaded end-to-end Convolutional Neural Network. IEEE Trans. Geosci. Remote Sens. 55, 3322–3337. https://doi.org/10.1109/TGRS.2017.2669341.

Colomina, I., Molina, P., 2014. Unmanned aerial systems for photogrammetry and remote sensing: A review. ISPRS J. Photogramm. Remote Sens. 92, 79–97. https://doi.org/10.1016/j.isprsjprs.2014.02.013.

Cortés, A.J., Garzón, L.N., Valencia, J.B., Madriñán, S., 2018. On the causes of rapid diversification in the páramos: Isolation by ecology and genomic divergence in espeletia. Front. Plant Sci. 871 https://doi.org/10.3389/fpls.2018.01700.

Cuatrecasas, J., 2013. A systematic study of the subtribe Espeletiinae. The New York Botanical Garden, New York, USA.

Deng, Z., Sun, H., Zhou, S., Zhao, J., Lei, L., Zou, H., 2018. Multi-scale object detection in remote sensing imagery with convolutional neural networks. ISPRS J. Photogramm. Remote Sens. 145, 3–22. https://doi.org/10.1016/j.isprsjprs.2018.04.003.

Diakogiannis, F.I., Waldner, F., Caccetta, P., Wu, C., 2020. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. ISPRS J. Photogramm. Remote Sens. 162, 94–114. https://doi.org/10.1016/j.isprsjprs.2020.01.013.

Diazgranados, M., 2017. Apuntes para la Revisión del Estado de Conservación y Amenaza de los Frailejones en Colombia. In: Moreno Gaona, D.A. (Ed.), IX Congreso Colombiano de Botánica. Ciencia en Desarrollo, Tunja, Boyacá (Colombia), p. 250.

Diazgranados, M., 2012. A nomenclator for the frailejones (Espeletiinae Cuatrec., Asteraceae). PhytoKeys 16, 1–52. https://doi.org/10.3897/phytokeys.16.3186.

Diazgranados, M., Barber, J.C., 2017. Geography shapes the phylogeny of frailejones (Espeletiinae Cuatrec., Asteraceae): A remarkable example of recent rapid radiation in sky islands. PeerJ 2017. https://doi.org/10.7717/peerj.2968.

Diazgranados, M., Castellanos, C., 2020. Libro Rojo de Frailejones de Colombia. Instituto de Investigación de Recursos Biológicos Alexander von Humboldt Volumen 2.

Falk, T., Mai, D., Bensch, R., Çiçek, Ö., Abdulkadir, A., Marrakchi, Y., Böhm, A., Deubner, J., Jäckel, Z., Seiwald, K., Dovzhenko, A., Tietz, O., Dal Bosco, C., Walsh, S., Saltukoglu, D., Tay, T.L., Prinz, M., Palme, K., Simons, M., Diester, I., Brox, T., Ronneberger, O., 2019. U-Net: deep learning for cell counting, detection, and morphometry. Nat. Methods 16, 67–70. https://doi.org/10.1038/s41592-018-0261-2.

Feng, W., Sui, H., Huang, W., Xu, C., An, K., 2019. Water Body Extraction from Very High-Resolution Remote Sensing Imagery Using Deep U-Net and a Superpixel-Based Conditional Random Field Model. IEEE Geosci. Remote Sens. Lett. 16, 618–622. https://doi.org/10.1109/LGRS.2018.2879492.

Fu, G., Liu, C., Zhou, R., Sun, T., Zhang, Q., 2017. Classification for High Resolution Remote Sensing Imagery Using a Fully Convolutional Network. Remote Sens. 9, 498. https://doi.org/10.3390/rs9050498.

Fu, K., Chang, Z., Zhang, Y., Xu, G., Zhang, K., Sun, X., 2020. Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images. ISPRS J. Photogramm. Remote Sens. 161, 294–308. https://doi.org/10.1016/j.isprsjprs.2020.01.025.

Gao, L., Song, W., Dai, J., Chen, Y., 2019. Road extraction from high-resolution remote sensing imagery using refined deep residual convolutional neural network. Remote Sens. 11, 1–16. https://doi.org/10.3390/rs11050552.

Geng, J., Jiang, W., Deng, X., 2020. Multi-scale deep feature learning network with bilateral filtering for SAR image classification. ISPRS J. Photogramm. Remote Sens. 167, 201–213. https://doi.org/10.1016/j.isprsjprs.2020.07.007.

Graham, L.J., Spake, R., Gillings, S., Watts, K., Eigenbrod, F., 2019. Incorporating fine-scale environmental heterogeneity into broad-extent models. Methods Ecol. Evol. 10, 767–778. https://doi.org/10.1111/2041-210X.13177.

Gupta, A., Byrne, J., Moloney, D., Watson, S., Yin, H., 2020. Tree Annotations in LiDAR Data Using Point Densities and Convolutional Neural Networks. IEEE Trans. Geosci. Remote Sens. 58, 971–981. https://doi.org/10.1109/TGRS.2019.2942201.

Hamylton, S.M., Morris, R.H., Carvalho, R.C., Roder, N., Barlow, P., Mills, K., Wang, L., 2020. Evaluating techniques for mapping island vegetation from unmanned aerial vehicle (UAV) images: Pixel classification, visual interpretation and machine learning approaches. Int. J. Appl. Earth Obs. Geoinf. 89, 102085 https://doi.org/10.1016/j.jag.2020.102085.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.

He, N., Paoletti, M.E., Haut, J.M., Fang, L., Li, S., Plaza, A., Plaza, J., 2019. Feature extraction with multiscale covariance maps for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 57, 755–769. https://doi.org/10.1109/TGRS.2018.2860464.

Huang, B., Lu, K., Audebert, N., Khalel, A., Tarabalka, Y., Malof, J., Boulch, A., Saux, B. L., Collins, L., Bradbury, K., Lefèvre, S., El-Saban, M., 2018. Large-scale semantic classification: Outcome of the first year of inria aerial image labeling benchmark. In: International Geoscience and Remote Sensing Symposium (IGARSS), pp. 6947–6950. https://doi.org/10.1109/IGARSS.2018.8518525.

Kemker, R., Salvaggio, C., Kanan, C., 2018. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. ISPRS J. Photogramm. Remote Sens. 145, 60–77. https://doi.org/10.1016/j.isprsjprs.2018.04.014.

Kraaijenbrink, P.D.A., Shea, J.M., Pellicciotti, F., Jong, S.M.D., Immerzeel, W.W., 2016. Object-based analysis of unmanned aerial vehicle imagery to map and characterise surface features on a debris-covered glacier. Remote Sens. Environ. 186, 581–595. https://doi.org/10.1016/j.rse.2016.09.013.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep Convolutional Neural Networks. In: NIPS2012: Neural Information Processing Systems. Lake Tahoe, Nevada, pp. 1–9.

Li, Q., Mou, L., Liu, Q., Wang, Y., Zhu, X.X., 2018. HSF-Net: Multiscale deep feature embedding for ship detection in optical remote sensing imagery. IEEE Trans. Geosci. Remote Sens. 56, 7147–7161. https://doi.org/10.1109/TGRS.2018.2848901.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, in. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3431–3440. https://doi.org/10.1109/CVPR.2015.7298965.

Lv, X., Ming, D., Lu, T., Zhou, K., Wang, M., Bao, H., 2018. A new method for region-based majority voting CNNs for very high resolution image classification. Remote Sens. 10, 1–24. https://doi.org/10.3390/rs10121946.

Marmanis, D., Schindler, K., Wegner, J.D., Galliani, S., Datcu, M., Stilla, U., 2018. Classification with an edge: Improving semantic image segmentation with boundary detection. ISPRS J. Photogramm. Remote Sens. 135, 158–172. https://doi.org/10.1016/j.isprsjprs.2017.11.009.

Masiero, A., Fissore, F., Vettore, A., 2017. A low cost UWB based solution for direct georeferencing UAV photogrammetry. Remote Sens. 9 https://doi.org/10.3390/rs9050414.

Milas, A.S., Arend, K., Mayer, C., Simonson, M.A., Mackey, S., 2017. Different colours of shadows: classification of UAV images. Int. J. Remote Sens. 38, 3084–3100. https://doi.org/10.1080/01431161.2016.1274449.

Padilla-González, G.F., Diazgranados, M., Da Costa, F.B., 2017. Biogeography shaped the metabolome of the genus Espeletia: A phytochemical perspective on an Andean adaptive radiation. Sci. Rep. 7 https://doi.org/10.1038/s41598-017-09431-7.

O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2015, pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.

Sun, G., Huang, H., Zhang, A., Li, F., Zhao, H., Fu, H., 2019. Fusion of multiscale convolutional neural networks for building extraction in very high-resolution images. Remote Sens. 11 https://doi.org/10.3390/rs11030227.

Varela, A., Fuentes, L., Martínez, C., Medina, M., Jácome, J., 2017. Programa Nacional Evaluación del Estado y Afectación de los Frailejones en los Páramos de los Andes del Norte: Avances. In: Moreno Gaona, D.A. (Ed.), IX Congreso Colombiano de Botánica. Tunja, Boyacá (Colombia), pp. 244–245.

Volpi, M., Tuia, D., 2017. Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. IEEE Trans. Geosci. Remote Sens. 55, 881–893. https://doi.org/10.1109/TGRS.2016.2616585.

Woellner, R., Wagner, T.C., 2019. Saving species, time and money: Application of unmanned aerial vehicles (UAVs) for monitoring of an endangered alpine river specialist in a small nature reserve. Biol. Conserv. 233, 162–175. https://doi.org/10.1016/j.biocon.2019.02.037.

Yang, Z., Dong Mu, X., Zhao, F., 2018. Scene classification of remote sensing image based on deep network and multi-scale features fusion. Optik (Stuttg). 171, 287–293. https://doi.org/10.1016/j.ijleo.2018.06.024.

Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., Xu, H., Tan, W., Yang, Q., Wang, J., Gao, J., Zhang, L., 2020. Deep learning in environmental remote sensing: Achievements and challenges. Remote Sens. Environ. 241, 1–24. https://doi.org/10.1016/j.rse.2020.111716.

Yue, K., Yang, L., Li, R., Hu, W., Zhang, F., Li, W., 2019. TreeUNet: Adaptive Tree convolutional neural networks for subdecimeter aerial image segmentation. ISPRS J. Photogramm. Remote Sens. 156, 1–13. https://doi.org/10.1016/j.isprsjprs.2019.07.007.

Zhang, C., Atkinson, P.M., 2016. Novel shape indices for vector landscape pattern analysis. Int. J. Geogr. Inf. Sci. 30, 2442–2461. https://doi.org/10.1080/13658816.2016.1179313.

Zhang, C., Harrison, P.A., Pan, X., Li, H., Sargent, I., 2020. Scale Sequence Joint Deep Learning (SS-JDL) for land use and land cover classi fi cation. Remote Sens. Environ. 237, 111593 https://doi.org/10.1016/j.rse.2019.111593.

Zhang, Chunju, Li, G., Du, S., 2019a. Multi-Scale Dense Networks for Hyperspectral Remote Sensing Image Classification. IEEE Trans. Geosci. Remote Sens. 57, 9201–9222. https://doi.org/10.1109/TGRS.2019.2925615.

Zhang, C., Pan, X., Li, H., Gardiner, A., Sargent, I., Hare, J., Atkinson, P.M., 2018a. A hybrid MLP-CNN classifier for very fine resolution remotely sensed image classification. ISPRS J. Photogramm. Remote Sens. 140, 133–144. https://doi.org/10.1016/j.isprsjprs.2017.07.014.

Zhang, C., Sargent, I., Pan, X., Gardiner, A., Hare, J., Atkinson, P.M., 2018b. VPRS-Based regional decision fusion of CNN and MRF classifications for very fine resolution remotely sensed images. IEEE Trans. Geosci. Remote Sens. 56, 4507–4521. https://doi.org/10.1109/TGRS.2018.2822783.

Zhang, Ce, Sargent, I., Pan, X., Li, H., Gardiner, A., Hare, J., Atkinson, P.M., 2019b. Joint Deep Learning for land cover and land use classification. Remote Sens. Environ. 221, 173–187. https://doi.org/10.1016/j.rse.2018.11.014.

Zhang, C., Sargent, I., Pan, X., Li, H., Gardiner, A., Hare, J., Atkinson, P.M., 2018c. An object-based convolutional neural network (OCNN) for urban land use classification. Remote Sens. Environ. 216, 57–70. https://doi.org/10.1016/j.rse.2018.06.034.

Zhang, W., Song, K., Rong, X., Li, Y., 2019c. Coarse-to-Fine UAV Target Tracking with Deep Reinforcement Learning. IEEE Trans. Autom. Sci. Eng. 16, 1522–1530. https://doi.org/10.1109/TASE.2018.2877499.

Zhang, Z., Liu, Q., Wang, Y., 2018d. Road Extraction by Deep Residual U-Net. IEEE Geosci. Remote Sens. Lett. 15, 749–753. https://doi.org/10.1109/LGRS.2018.2802944.

Zou, Q., Ni, L., Zhang, T., Wang, Q., 2015. Deep Learning Based Feature Selection for Remote Sensing Scene Classification. IEEE Geosci. Remote Sens. Lett. 12, 2321–2325. https://doi.org/10.1109/LGRS.2015.2475299.