remote sensing

MDPI

Article

Transfer Learning-Based Accurate Detection of Shrub Crown Boundaries Using UAS Imagery

Jiawei Li, Huihui Zhang * and David Barnard

Water Management and Systems Research Unit, United States Department of Agriculture, Agricultural Research Service, 2150 Centre Avenue, Building D, Fort Collins, CO 80526, USA; jiawei.li@usda.gov (J.L.); david.barnard@usda.gov (D.B.)

* Correspondence: huihui.zhang@usda.gov

Abstract

The accurate delineation of shrub crown boundaries is critical for ecological monitoring, land management, and understanding vegetation dynamics in fragile ecosystems such as semi-arid shrublands. While traditional image processing techniques often struggle with overlapping canopies, deep learning methods, such as convolutional neural networks (CNNs), offer promising solutions for precise segmentation. This study employed high-resolution imagery captured by unmanned aircraft systems (UASs) throughout the shrub growing season and explored the effectiveness of transfer learning for both semantic segmentation (Attention U-Net) and instance segmentation (Mask R-CNN). It utilized pre-trained model weights from two previous studies that originally focused on tree crown delineation to improve shrub crown segmentation in non-forested areas. Results showed that transfer learning alone did not achieve satisfactory performance due to differences in object characteristics and environmental conditions. However, fine-tuning the pre-trained models by unfreezing additional layers improved segmentation accuracy by around 30%. Fine-tuned pre-trained models show limited sensitivity to shrubs in the early growing season (April to June) and improved performance when shrub crowns become more spectrally unique in late summer (July to September). These findings highlight the value of combining pre-trained models with targeted fine-tuning to enhance model adaptability in complex remote sensing environments. The proposed framework demonstrates a scalable solution for ecological monitoring in data-scarce regions, supporting informed land management decisions and advancing the use of deep learning for long-term environmental monitoring.

Keywords: deep learning; shrub segmentation; UAS; attention U-Net; Mask R-CNN

check for updates

Academic Editor: Kevin Tansey

Received: 15 April 2025
Revised: 2 June 2025
Accepted: 27 June 2025
Published: 3 July 2025

Citation: Li, J.; Zhang, H.; Barnard, D. Transfer Learning-Based Accurate Detection of Shrub Crown Boundaries Using UAS Imagery. *Remote Sens.* **2025**, *17*, 2275. https://doi.org/10.3390/rs17132275

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Semi-arid shrublands, which cover over 18% of the Earth's terrestrial surface [1,2], play a crucial role in global carbon cycling, biodiversity conservation, and soil stabilization. However, these ecosystems are highly vulnerable to climate change and desertification, making their monitoring essential for sustainable land management [3]. Effective monitoring requires the accurate delineation of shrub crown boundaries, which is vital for assessing vegetation dynamics, ecosystem health, and conservation strategies. Despite the importance of semi-arid shrublands, mapping them in these regions presents significant challenges for traditional remote sensing methods. Techniques such as object-based image analysis (OBIA) and edge detection algorithms often struggle to detect shrubs due to their small size and spectral similarity to surrounding grasses or low-lying plants [4]. These

challenges are further exacerbated by seasonal spectral variations and limited annotated training data, making precise shrub delineation difficult [5].

Recent advances in artificial intelligence (AI) have transformed remote sensing by improving feature extraction and segmentation accuracy in high-resolution imagery [6]. Unlike traditional machine learning, which relies on significant feature engineering and preprocessing for unstructured data (e.g., images, sound), deep learning methods like convolutional neural networks (CNNs) can automatically learn hierarchical features, improving object recognition in complex landscapes. Several studies have considered deep learning as an alternative to OBIA for classification and object detection. Guirado et al. [7] utilized both OBIA and CNN methods for the object detection of the *Ziziphus lotus* shrub, achieving an average of a 95% F1-score and showing that the best CNN detector achieved up to 12% better precision, up to 30% better recall, and up to 20% better balance between precision and recall than the OBIA method. James & Bradshaw [8] demonstrated the feasibility of using deep learning and drone technology for real-time plant species detection in a farm southwest of Grahamstown, in the Eastern Cape, South Africa, achieving 83% F1 in detecting shrubs, highlighting the potential of using deep learning for efficient and scalable vegetation monitoring. Khaldi et al. [9] presented a deep learning approach for the individual mapping of large polymorphic shrubs in high-mountain ecosystems in the Sierra Nevada National Park located on the southern fringe of the Iberian Peninsula using satellite imagery, and achieved an F1-score in shrub delineation of 87.87% on the photo-interpreted data, demonstrating deep learning's effectiveness for regional-scale medium-to-large shrub detection using high-resolution google earth satellite imagery (13 cm resolution).

Despite the notable results achieved with deep learning in shrub detection, training deep learning models typically requires large, annotated datasets [10]. This requirement is often impractical in remote sensing applications, where field data collection and manual labeling are labor-intensive and resource-demanding. Consequently, there is a growing need for methodologies that enable robust model performance using limited training data. Moreover, in the broader field of computer vision, small object detection remains a well-known challenge due to the limited number of pixels representing small objects, making them more susceptible to misclassification or being overlooked [11]. This issue is particularly evident in shrublands, where shrubs only occupy a small portion of high-altitude unmanned aerial system (UAS) imagery, further complicating detection. The framework for accurately delineating shrubs remains challenging, as current methods face limitations in handling variability in shrub morphology, density, and environmental conditions.

To address the challenges posed by complex environmental conditions and limited training data, transfer learning and data augmentation have proven effective in alleviating data constraints [12–14]. Specifically, transfer learning reduces the reliance on large, labeled datasets by transferring knowledge from pre-trained models, while data augmentation improves model robustness, prevents overfitting, and addresses the issue of limited data by generating new synthetic data points through various transformations of the original dataset. This study explored the potential of using pre-trained models that were initially trained for tree crown delineation [15,16] in shrub detection. Given the shared structural traits between trees and shrubs, such as the presence of canopies and vegetative textures, these models offer a potential starting point for accurate shrub delineation. However, a key challenge lies in the domain shift between the source domain (trees) and the target domain (shrubs). Domain shift refers to the difference in data distributions between the domains, which can degrade model performance when the learned features in the source domain do not generalize well to the target [17]. In this context, morphological differences are particularly relevant: trees often exhibit larger, more distinct, and vertically prominent crowns, while shrubs have smaller, irregular, and more horizontally distributed

canopies that often overlap and blend with background elements such as soil or grass. These structural differences affect spatial, spectral, and contextual features, potentially reducing the effectiveness of the pre-trained representations. To address this, our study employed fine-tuning to refine these models. Fine-tuning with labeled shrubland data helps the model recalibrate its feature detectors to focus on shrub-specific traits, such as irregular crown structures, overlapping canopies, and increased background noise [18]. This targeted adaptation is essential for overcoming domain shifts and achieving accurate shrub delineation in complex environments.

Although previous research has been conducted on shrub detection, much of it has focused on local-scale applications with extensive effort in data annotation and significant computational resources. The objective of this study was to test whether transfer learning can enhance the generalizability and efficiency of shrub crown detection by reducing annotation and computational demands while maintaining competitive accuracy. To address this objective, we investigated whether pre-trained models from related domains can be effectively adapted to shrub delineation tasks through fine-tuning, by evaluating the performance of transferred models based on two fine-tuning levels. The first level involves feature extraction, where only the final classification layers are retrained, while the second level applies more extensive fine-tuning, with more layers unfrozen for retraining.

2. Materials and Methods

2.1. Background

2.1.1. Study Area

This research was conducted in a shrubland field near Severance, Colorado, USA, a region characterized as semi-arid with low annual precipitation (around 330 mm), as shown in Figure 1a. This area is part of the North American shortgrass prairie ecoregion, where shrub encroachment and climate-driven land degradation pose significant threats to ecosystem stability [19]. This field features a relative upland topographic position with a relief of 29 m within the current 109 ha field. The area was initially planted with native grass and shrubs under the Farm Service Agency's Conservation Reserve Program (CRP) around 1988, then tilled again in the 1990s, creating 12 alternating strips of wheat/fallow, each with a new declination of approximately 12° east of true north and a width of about 120 m. The field was then converted to CRP again with seeding treatments applied to alternating rows in 2013 and 2014. These two years had drastically different drought conditions at the time of seeding, resulting in contrasting vegetation communities despite using identical seed mixes [20]. Hence, the heterogeneous landscape that emerged includes distinct zones, with a high abundance of *Atriplex canescens* (Fourwing saltbush) alternating with rows containing primarily short grass, forbs, bare soil, and fewer shrubs. The fragmented distribution of shrub crowns and the spectral similarity between sensed vegetation and short grass make this region a representative case study for testing remote sensing methods in extreme, data-scarce environments. The columnar planting pattern also provides a controlled spatial framework to assess segmentation accuracy across varying shrub densities.

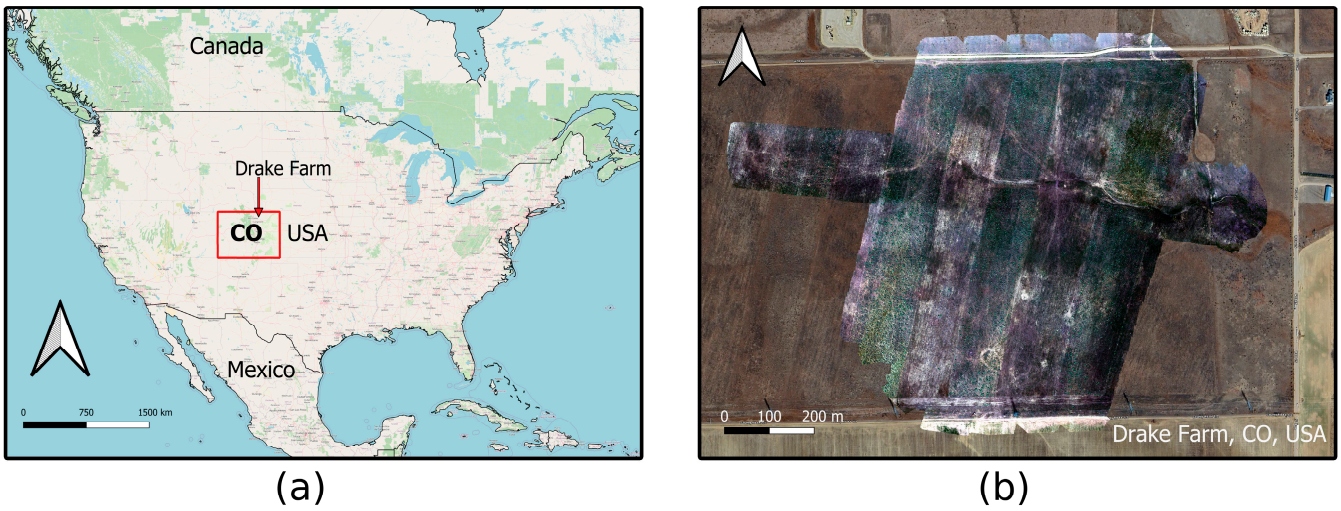


Figure 1. (**a**) The location of the Drake farm in Colorado, USA. (**b**) An example of UAS orthomosaic RGB imagery of Drake Farm, taken on 28 September 2022.

2.1.2. UAS Data Acquisition

This study utilized high-resolution multispectral imagery captured by a DJI Spreading Wings S900 hexacopter (DJI Co., Ltd., Shenzhen, China) equipped with a Micasense RedEdge-MX multispectral camera (MicaSense Inc., Seattle, WA, USA) under sunny and calm weather conditions, with wind speeds ranging from 2 to 7 m/s. The multispectral camera captured data in five bands: Blue (475 ± 32 nm), Green (560 ± 27 nm), Red (668 ± 16 nm), RedEdge (717 nm ± 12 nm), and NIR (842 ± 57 nm). Flights were conducted at a height of 120 m above ground level (AGL) with an overlap/sidelap of 88/70%, resulting in a spatial resolution of 0.03 m. Image processing was performed using Agisoft Metashape software (Agisoft LLC, Saint Petersburg, Russia) to create orthomosaics and georectify the data. Imagery was collected biweekly from April 2022 to September 2022, with a total of 10 flights. The Micasense RP04-1826289-SC calibration plate was used for image calibration. The survey covers 1285 km^2, with an estimated 10,000 to 12,000 saltbushes grown in the survey region; an example of UAS orthomosaic RGB imagery is shown in Figure 1b.

2.1.3. Dataset Preparation

The UAS imagery was systematically divided into 1000 × 1000-pixel patches, with a 100-pixel overlap to enhance model performance and minimize boundary-related artifacts, resulting in approximately 169 image tiles per flight. The annotation of shrubs was based on the images taken on 28 September 2022. Following the annotated data fractions recommended by Sun et al. [21], a total of 30 image tiles were randomly selected, and 1500 individual shrub crowns were manually delineated in these selected 30 image tiles to create the project dataset, which is roughly 10% of the total shrub crowns on the farm. Each shrub mask was labeled based on distinguishable features, such as identifiable canopy shadows and textural differences (Figure 2). Although the training used only RGB imagery, the labeling process used the near-infrared (NIR) band for distinguishing live vegetation from bare soil. Adjoining crowns were carefully separated to ensure that each shrub was represented as a distinct segment to accurately capture their growth patterns and improve annotation precision. The dataset was then formatted in the Common Objects in Context (COCO) dataset structure to support efficient model training and evaluation [22].

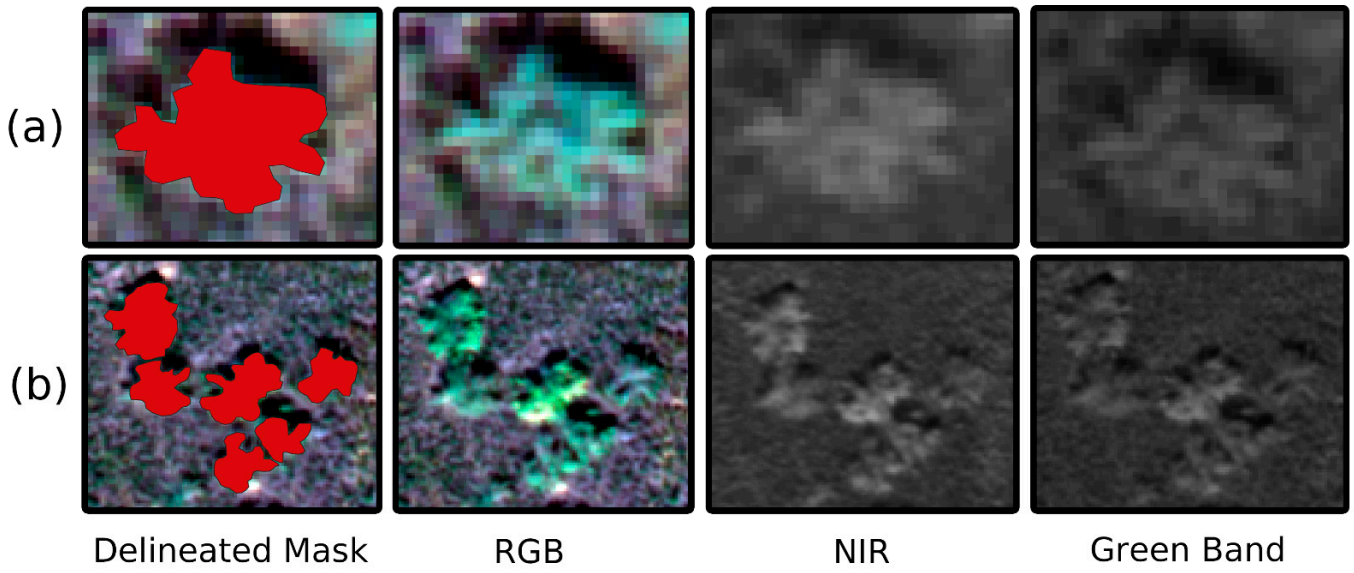


Figure 2. Examples of manual delineation of shrubs. Shrubs were labeled based on RGB, NIR, and green bands. (**a**) An example of an individual shrub. (**b**) An example of a dense shrub cluster.

For instance, for segmentation tasks in this study, bounding boxes were calculated based on the delineation mask boundaries of individual shrubs. The dataset was partitioned into distinct training, validation, and test sets to ensure unbiased performance evaluation. A total of 1200 shrubs were allocated for training, while 150 shrubs were designated for validation and 150 for testing. The validation and test datasets were isolated during the training process for independent assessment of the model's performance.

2.2. Methodology

2.2.1. Data Augmentation

To address the challenge of limited training data, both geometric and spectral augmentation strategies were employed to enhance model robustness and improve generalization using the Albumentations open-source package [23]. This augmentation strategy was designed to improve the model's ability to detect shrub crowns under variable conditions while mitigating the limitations of the small training dataset.

Geometric transformations such as flipping, rotation, and cropping simulated real-world variations in object position, scale, and orientation typical in UAS imagery, enhancing the model's adaptability to different viewpoints. Elastic deformations and affine transformations introduced controlled distortions from sensor errors or perspective shifts, helping the model to handle imperfect data. Spectral augmentations, including random changes to brightness, contrast, tone curves, and color balance, addressed variations in lighting and atmospheric conditions, making the model more resilient and accurate in diverse, real-world environments [24]. A complete specification of augmentation parameters is presented in Table A1.

2.2.2. Model Architecture

A comprehensive evaluation of the existing model architectures previously applied in vegetation detection [15,16] was used to assess the feasibility of transfer learning for shrub identification. Both semantic segmentation and instance segmentation approaches (described below) were explored to provide a well-rounded assessment of shrub detection performance across different computer vision tasks. Semantic segmentation was employed for pixel-level classification, which proved particularly useful for quantifying sparse vegetation in degraded control plots. Conversely, instance segmentation was applied to enable

individual crown delineation, offering potential insights for shrub species identification. Both models adapted from previous research were originally trained on RGB images. Therefore, despite multispectral data being collected in this study, the fine-tuning process used only the RGB bands of the dataset.

The semantic segmentation task was adapted from the research by Li et al. [15], which is an attention U-Net model focused on tree counting, crown segmentation, and height prediction. This study only adapted the tree counting and crown segmentation model described as "Model 1" in Li et al. [15]; the model weight was obtained from the author's GitHub repositories. Figure 3 shows the model diagram; detailed model parameters are in Table A2. The transfer learning process was implemented at two levels to evaluate its effectiveness. In the first level, only the output layer (L27B), attention layers (L15, L18, L21, and L24), and shallow decoder layers (L25 and L26) were unfrozen to allow the model to adjust high-level feature representations (e.g., canopy shape, size, etc.) without modifying the general feature extraction capabilities of the deeper layers (e.g., texture, shadow, etc.). If this approach did not produce satisfactory results, the second level would unfreeze the entire decoder layer (L15–L27B) and partially unfreeze the shallow encoder layers (L10–L14) for further fine-tuning, enabling the model to adjust more low-level features.

Prior to training, the input dataset was standardized and cropped to the required input size (256 × 256 pixels) before being fed into the network in batches of eight. In level one, the initial learning rate of the Adam optimizer [25] for the output layer (L27B) was set to 1×10^{-3}, the attention layers (L15, L18, L21, and L24) were set to 1×10^{-4}, the decoder layers were set to 1×10^{-4}, and the maximum training epochs were set to 30 epochs, with early stopping to avoid overfitting. In level two, the initial learning rate of the Adam optimizer for the decoder layer (L15–L26) was set to 1×10^{-4}, the shallow encoder layers (L10–L14) were set to 1×10^{-5}, the output layer (L27B) was set to 1×10^{-3}, and the maximum training epochs were set to 50 epochs. For evaluating the segmentation performance against manual delineations, and to be able to compare between segmentation tasks, the postprocessing started by extracting the contours from the segmentation mask using watershed segmentation. The IoU was calculated by comparing the predicted polygons with the ground truth polygons. The probability of the instance was calculated by averaging the probability inside the predicted instance. We then computed the object-wise AP (average precision) to compare with instance segmentation task results.

The instance segmentation task was adapted from Ball et al. [16], which is a Mask R-CNN architecture [26] with a ResNet-101 backbone [27]. It was attached with a Feature Pyramid Network (FPN) [28] and focused on accurately delineating individual tree crowns in tropical forests using aerial RGB imagery. The model architecture is briefly illustrated in Figure 4. The model weight was obtained from the author's GitHub repositories. The input dataset was standardized before being fed into the Mask R-CNN in batches of four. The transfer learning process was implemented in two levels to evaluate the effectiveness of fine-tuning. In level one, only the shallow stage (stage 4) of the backbone was unfrozen to allow the model to adjust high-level feature representations (e.g., canopy shape, size, etc.) without modifying the general feature extraction capabilities of the deeper stages (e.g., texture, shadow, etc.), and the maximum training epochs were set to 30 epochs, with early stopping to avoid overfitting. Level two would unfreeze deeper stages (stages 2, 3, and 4) in the backbone for further fine-tuning, enabling the model to adjust more low-level features, and the maximum training epochs were set to 50 epochs. Both levels use an SGD optimizer [29] with a momentum and WarmupMultiStepLR learn rate scheduler [30,31]. The parameters during training in both steps are listed in Tables 1–3.

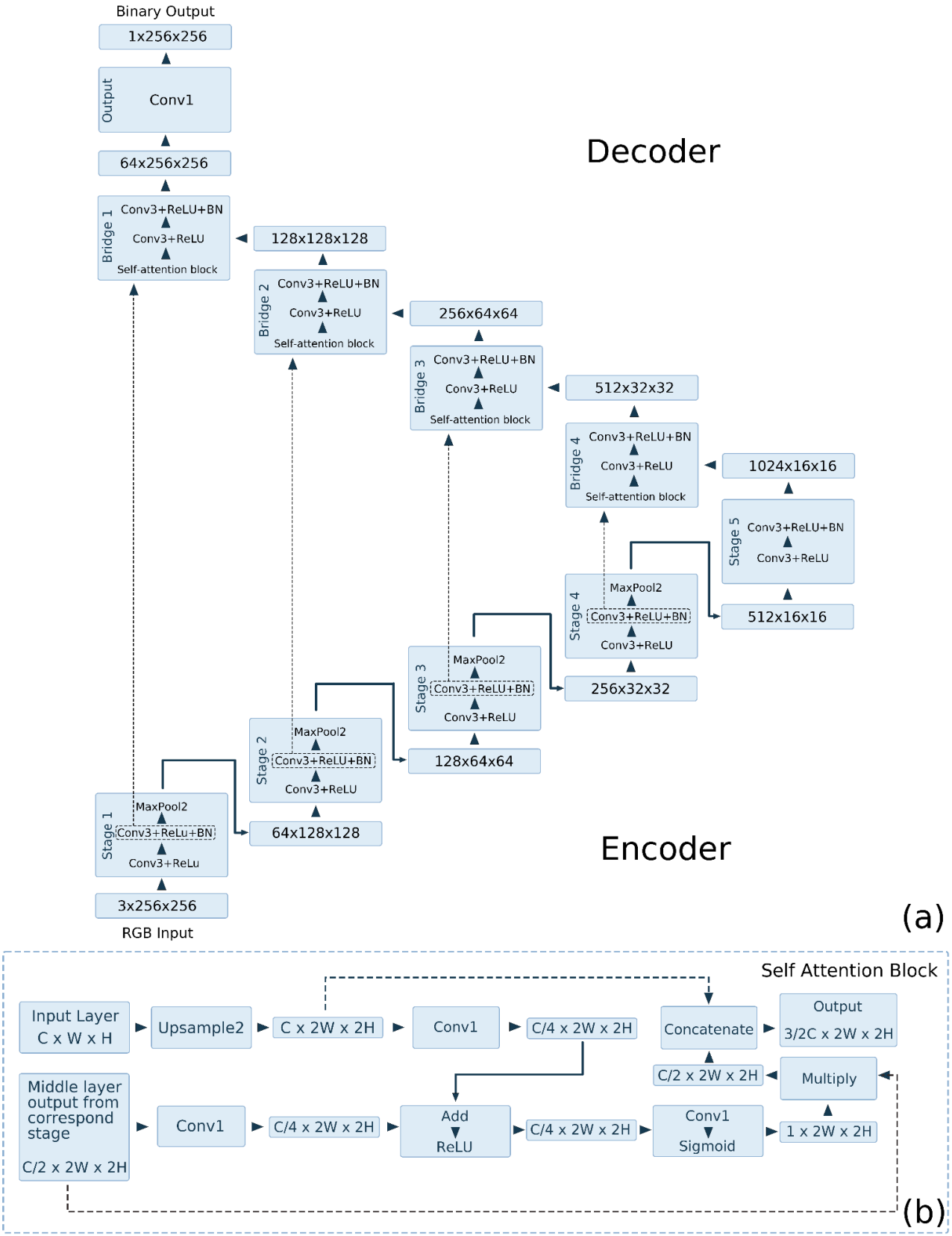


Figure 3. (**a**) Network architecture of Attention U-Net, featuring encoder–decoder stages with skip connections and self-attention blocks that refine feature fusion for improved segmentation accuracy. (**b**) The detailed self-attention block architecture.

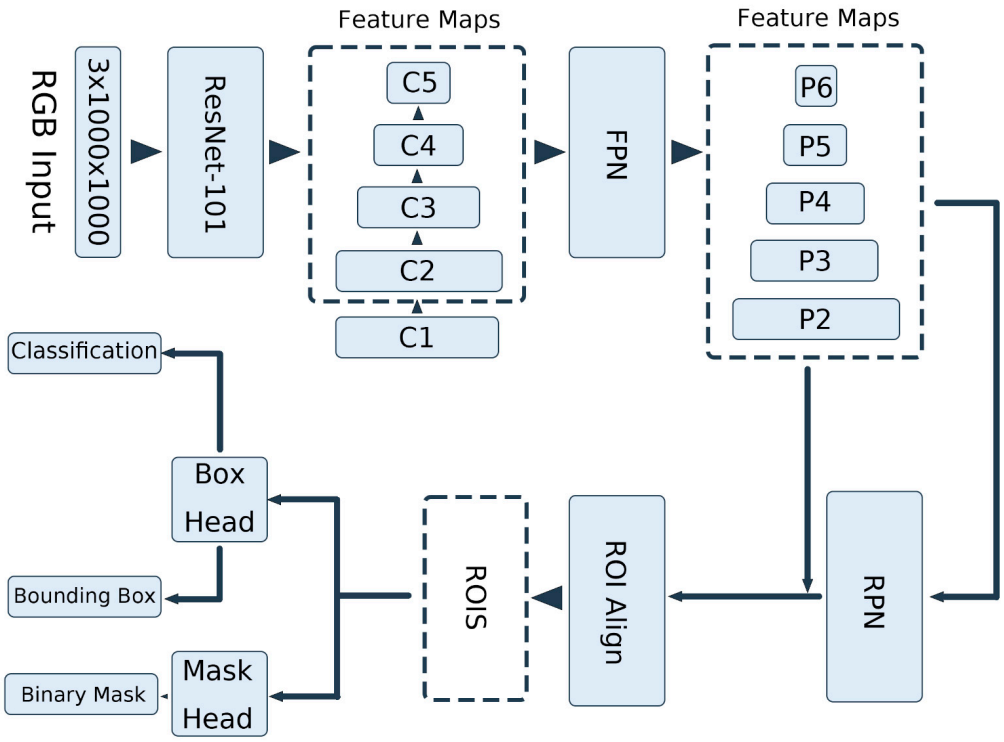


Figure 4. Network architecture of Mask R-CNN, featuring a ResNet-FPN backbone for multi-scale feature extraction, a Region Proposal Network (RPN) for generating object candidates, and ROIAlign modules followed by separate heads for object classification, bounding box regression, and mask prediction.

Table 1. The optimizer parameters during transfer learning in both steps.

	Base Learn Rate	Momentum	Weight Decay
Level one	1×10^{-3}	0.9	1×10^{-4}
Level two	5×10^{-3}	0.9	1×10^{-4}

Table 2. The learn rate scheduler parameters during transfer learning in both steps.

	Max Iteration	Warmup Iters	Warmup Factor	Steps	Gamma
Level one	5000	1000	1×10^{-3}	(2500, 3750)	0.1
Level two	10,000	1000	1×10^{-3}	(6000, 8000)	0.1

Table 3. The learn rate factor in Mask R-CNN for each block during training in both steps.

	Backbone	FPN	ROI Head
Level one	0.1	1.0	1.0
Level two	0.1	1.0	1.0

To evaluate model performance, the Mask R-CNN was assessed at the end of each learning rate cycle (every 4 epochs) by predicting shrub masks. The bounding boxes and category output were ignored. Performance was measured using average precision (AP) [32] to evaluate mask segmentation quality. The evaluation was conducted using the Pycocotools open-source package [22].

3. Results

3.1. Attention U-Net

The transfer learning results using a pre-trained model for attention U-Net, with only the output layer, attention layers, and shallow decoder layers unfrozen (described as level one in Section 2.2.2), and with the entire decoder layer and shallow encoder layers unfrozen (described as level two in Section 2.2.2) are presented in Table 4. The model fine-tuned under level two achieved better performance across all evaluation metrics. Specifically, the AP@0.5 increased from 0.28 (95% CI: 0.25–0.31) in level one to 0.60 (95% CI: 0.40–0.78) in level two. Similarly, AP@0.75 improved from 0.17 to 0.53, and the AP@[0.5:0.95] improved from 0.13 to 0.46, indicating enhanced detection accuracy and robustness with deeper fine-tuning.

Table 4. The transfer learning result for Attention U-Net.

Attention U-Net	AP@0.5	AP@0.75	AP@[0.5:0.95]
Level one	0.28 (95% CI: 0.25–0.31)	0.17 (95% CI: 0.13–0.21)	0.13 (95% CI: 0.08–0.18)
Level two	0.60 (95% CI: 0.40–0.78)	0.53 (95% CI: 0.30–0.75)	0.46 (95% CI: 0.26–0.66)

Examples of the prediction results are shown in Figure 5, including the ground truth and the outputs generated by the Attention U-Net model after each fine-tuning stage. The "Valid" column shows the ground truth masks overlaid in red. At level one, the model identifies many target regions, but with limited precision and over-segmentation in some areas. In contrast, level two produces predictions that are more consistent with the ground truth in terms of both shape and location, especially evident in regions with dense or irregularly shaped objects. These visual improvements correspond well with the quantitative gains observed in Table 4, demonstrating the benefit of unfreezing additional layers for transfer learning.

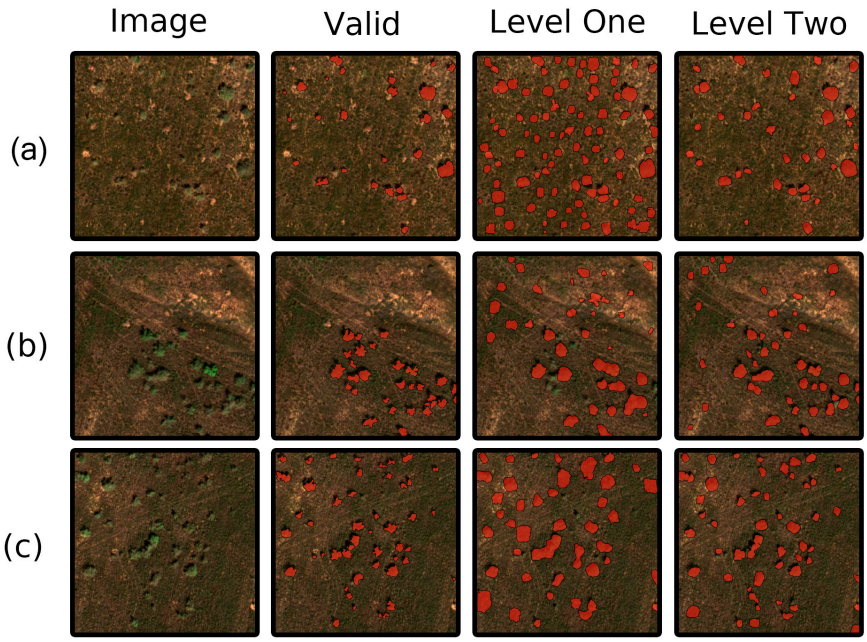


Figure 5. The comparison of two fine-tuning levels for Attention U-Net using the test dataset in the context of transfer learning. (**a**) Example area with relatively homogeneous shrub distribution and bare soil. (**b**) Example area with relatively heterogeneous shrub distribution and bare soil. (**c**) Example area with shrubs on the grassland.

3.2. Mask R-CNN

The transfer learning results using the Mask R-CNN model, with two levels of fine-tuning, are presented in Table 5. Similarly to the Attention U-Net experiment, level one refers to a limited fine-tuning setup, while level two involves the deeper unfreezing of the model's layers. The results indicate a notable improvement in detection performance when more layers are unfrozen. Specifically, AP@0.5 increased from 0.22 (95% CI: 0.09–0.34) in level one to 0.52 (95% CI: 0.42–0.60) in level two. Likewise, AP@0.75 improved from 0.14 to 0.38, and AP@[0.5:0.95] improved from 0.14 to 0.36, suggesting that more comprehensive fine-tuning leads to better generalization and object delineation across various IoU thresholds.

Table 5. The transfer learning results for Mask R-CNN.

Mask R-CNN	AP@0.5	AP@0.75	AP@[0.5:0.95]
Level one	0.22 (95% CI: 0.9–0.34)	0.14 (95% CI: 0.02–0.26)	0.14 (95% CI: 0.00–0.28)
Level two	0.52 (95% CI: 0.42–0.60)	0.38 (95% CI: 0.25–0.50)	0.36 (95% CI: 0.16–0.54)

A preview of the prediction results is shown in Figure 6, including the ground truth and the outputs generated by the Mask R-CNN model after each fine-tuning stage. In all three image examples (rows a–c), level one predictions tend to produce large, over-smoothed masks that frequently merge adjacent objects or misrepresent their actual boundaries. In contrast, the level two predictions are more refined and closely resemble the spatial distribution and shapes of the ground truth annotations, especially in denser regions or areas with variable object sizes.

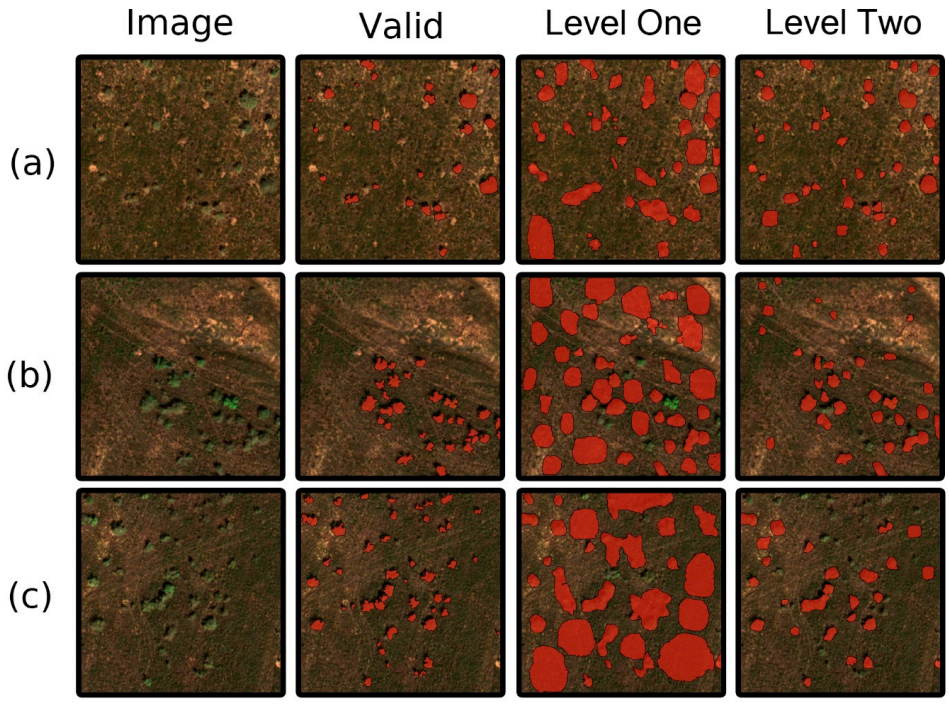


Figure 6. The comparison of two fine-tuning levels for Mask R-CNN using the test dataset in the context of transfer learning. (**a**) Example area with relatively homogeneous shrub distribution and bare soil. (**b**) Example area with relatively heterogeneous shrub distribution and bare soil. (**c**) Example area with shrubs on the grassland.

3.3. The Comparison of Fine-Tuning Results Between Two Pre-Trained Models

To evaluate and compare the performance of two pre-trained models in detecting shrubs of varying sizes, predictions were analyzed by stratifying the test set based on the shrub canopy area. Bootstrapping [33] was employed to estimate 95% confidence intervals for detection performance metrics via resampling with replacement. For each 1 m^2 canopy area bin, AP@0.5, AP@0.75, and AP@[0.5:0.95] were computed, and the bootstrapping results were plotted side by side as boxplots for comparison, as shown in Figure 7.

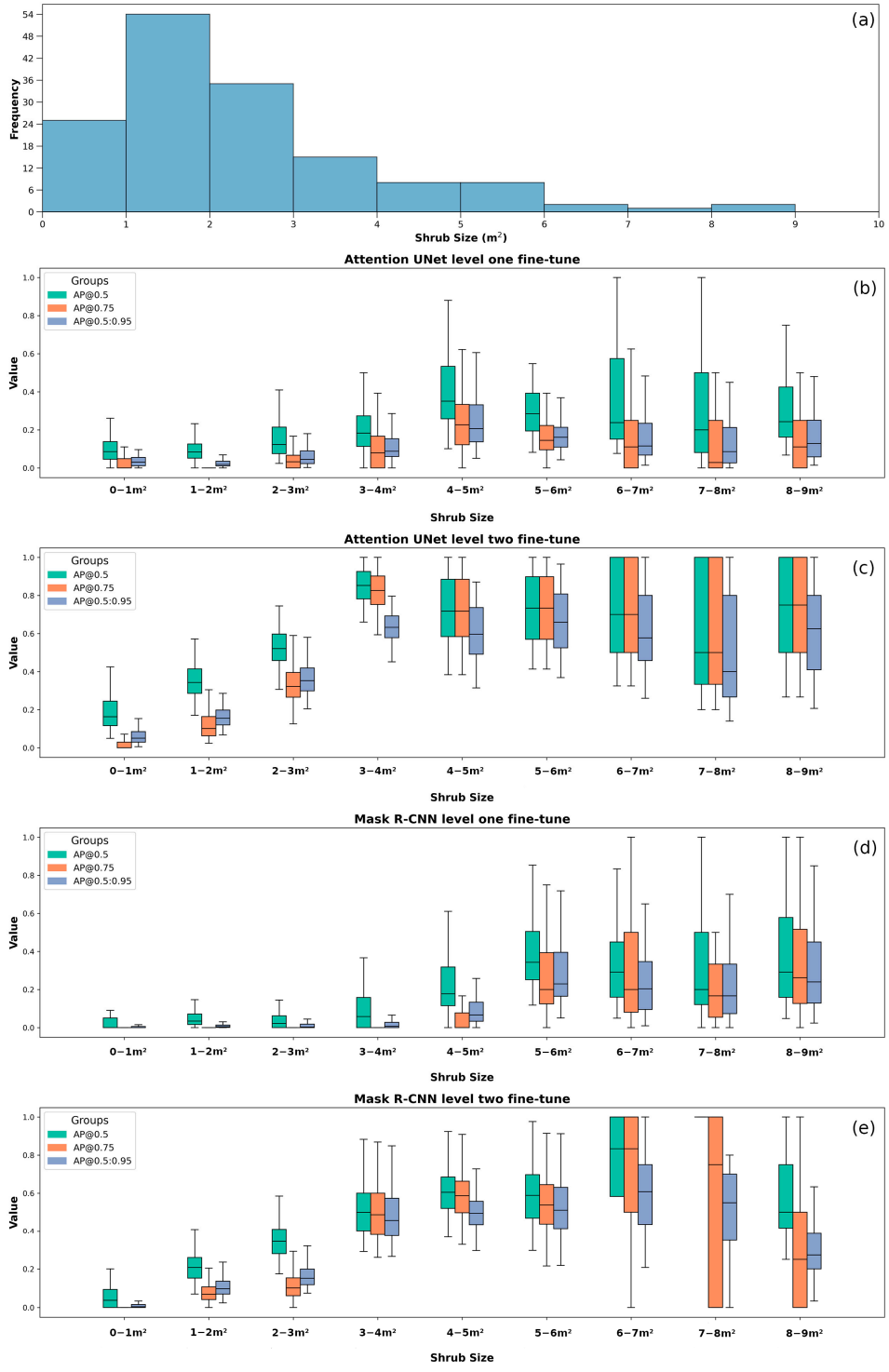


Figure 7. (**a**) The shrub pixel size distribution of the test set. (**b**) The AP@0.5, AP@0.75, and AP@[0.5:0.95] scores for each range of shrub size after level one fine-tuning of Attention U-Net. (**c**) The AP@0.5, AP@0.75,

and AP@[0.5:0.95] scores for each range of shrub size after level two fine-tuning of Attention U-Net. (**d**) The AP@0.5, AP@0.75, and AP@[0.5:0.95] scores for each range of shrub size after level one fine-tuning of Mask R-CNN. (**e**) The AP@0.5, AP@0.75, and AP@[0.5:0.95] scores for each range of shrub size after level two fine-tuning of Mask R-CNN.

Figure 7a shows the distribution of shrub sizes in the dataset, with the majority falling within the 1–3 m^2 range. Figure 7b,c illustrate the performance for level one and level two fine-tuning for the Attention U-Net model, and Figure 7c,d illustrate the performance for level one and level two fine-tuning for the Mask R-CNN model. During level one fine-tuning (Figure 7b), Attention U-Net exhibited limited performance for detecting small shrubs (<3 m^2), with all AP scores remaining close to zero in the first few bins. As the shrub size increased beyond 3 m^2, detection improved gradually, with AP@0.5 values reaching around 0.4–0.6 in the 5–8 m^2 range. After level two fine-tuning (Figure 7c), the model showed marked improvements, particularly for shrubs larger than 3 m^2. The AP@0.5 scores consistently reached above 0.6 in the 4–8 m^2 range, and the AP@0.75 and AP@[0.5:0.95] also improved. However, detection remained unreliable for smaller shrubs (<3 m^2), where all three AP metrics remained below 0.2, highlighting a continued challenge in capturing small object instances even with deeper fine-tuning.

Mask R-CNN followed a similar trend for level one fine-tuning. The AP values were low across the smallest size bins and improved slightly in the 4–6 m^2 range, though often with broader confidence intervals and more variability than Attention U-Net. The detection of small shrubs remained particularly weak, with negligible AP scores for shrubs smaller than 3 m^2. With level two fine-tuning (Figure 7e), Mask R-CNN showed improvements for medium-to-large shrubs ($\geq$4 m^2), with AP@0.5 values reaching up to 0.7 in some bins. However, results were inconsistent, and in some larger shrub size categories (e.g., 7–9 m^2), confidence intervals were wide and median AP scores dropped. Importantly, like Attention U-Net, performance on shrubs smaller than 3 m^2 remained very low across all three AP metrics, indicating that deeper fine-tuning alone did not resolve the limitations in detecting smaller vegetation patches.

3.4. The Prediction Results Through the Entire Growing Season

Figure 8 illustrates the temporal dynamics of shrub crown delineation using two fine-tuned deep learning models—Attention UNet and Mask R-CNN—across ten time points during the 2024 shrub growing season, ranging from early spring (26 April 2024) to early fall (28 September 2024). Each time point includes three visual representations: the original RGB image (bottom row of each time block), and the corresponding binary shrub segmentation masks predicted by the Attention UNet (top row) and Mask R-CNN (middle row).

As the season progresses, clear temporal trends in vegetation growth and canopy expansion are evident in the RGB imagery. Early-season images (e.g., April to early June) display minimal green coverage, reflecting limited shrub development. Correspondingly, both models detect fewer and smaller shrub crowns during this period. From mid-June through September, vegetation visibly becomes denser and more uniformly distributed in the RGB imagery, indicating peak shrub growth. This is reflected in the segmentation masks, where both models predict an increased number and size of shrub crowns.

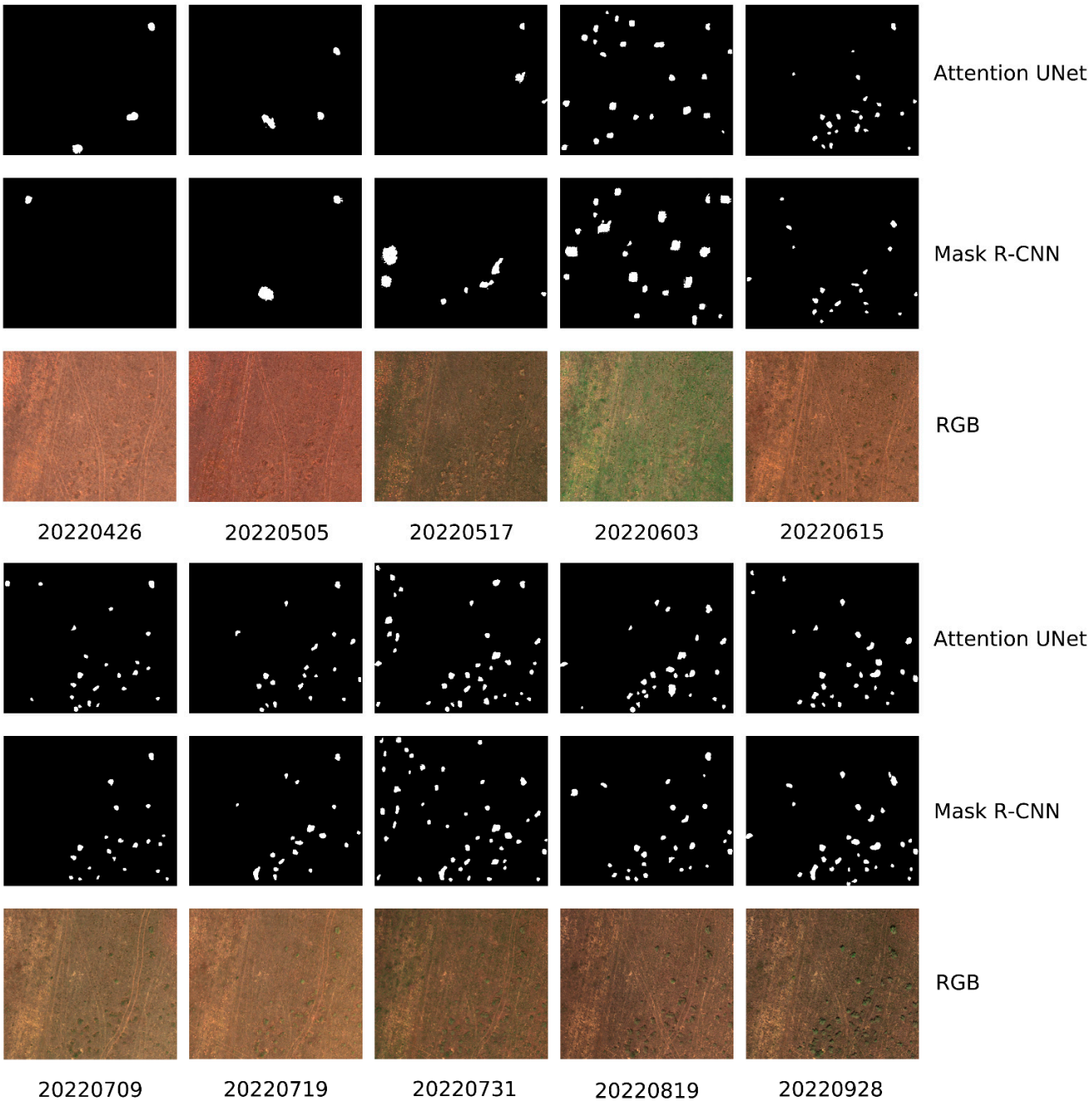


Figure 8. Example of shrub crown boundary detection across the shrub growing season using both level two fine-tuned Attention U-Net and Mask R-CNN models.

4. Discussion

4.1. Improved Shrub Detection Through Fine-Tuning

The Attention U-Net prediction results, shown in Table 4 and Figure 5 after level one fine-tuning, exhibit low average precision and a high number of false positives and false negatives. The model also struggles to distinguish shrubs that are spatially close to each other. This is likely due to the limited fine-tuning applied, where only the output layer, selected attention layers, and shallow decoder layers were unfrozen. With most encoder layers remaining fixed, the model lacks the flexibility to adapt the low- and mid-level features critical for accurate boundary detection. The model's performance after level one fine-tuning is essentially unusable for practical purposes. After level two fine-tuning, the prediction results show a substantial improvement, as detailed in Table 4. The enhanced

model performance is also visually evident in Figure 5, where many of the false positives and false negatives observed in earlier predictions have been significantly reduced. The models now demonstrate a more accurate ability to delineate shrub crowns, especially in areas with high shrub density or overlapping canopies. Notably, the model after level two fine-tuning is better equipped to distinguish and separate spatially adjacent shrubs, reducing instances of crown merging and improving the overall precision of individual crown detection. This improvement suggests that the additional training has enabled the model to achieve more reliable segmentation outcomes.

The Mask R-CNN prediction results after the first level of fine-tuning, shown in Table 4 and Figure 6, indicate a performance level that is also essentially unusable for practical purposes. The model exhibits an even lower average precision and an even higher number of false positives and false negatives. This poor performance is likely due to the very limited fine-tuning at this stage, where only a small subset of higher-level layers are unfrozen, preventing the model from effectively adapting its learned features to the new domain. Level two fine-tuning did improve the overall model performance, as shown in Table 5, and with better spatial delineation than level one fine-tuning, as shown in Figure 6. These visual results align well with the quantitative gains shown in Table 5 and demonstrate that deeper fine-tuning enables the model to better capture the fine-grained structure of the target objects. However, the Mask R-CNN model results in lower AP scores across all thresholds, and overall, more false negatives compared to with Attention U-Net.

The improved model performance of level two fine-tuning compared for both models observed in this study reinforces the value of fine-tuning when adapting pre-trained models to datasets with distinct characteristics. This approach is particularly relevant in remote sensing applications where environmental factors such as vegetation density, lighting conditions, and sensor variability can vary significantly between datasets, supported by previous research [34,35]. This adaptation of pre-trained models allows for more accurate and context-specific results, enabling the application of deep learning methods in natural resources, such as the monitoring of biodiversity, land cover, fuel structure, and land degradation, as highlighted by past research on shrub detection [36]. However, this study found that the Mask R-CNN model underperformed in comparison to Attention U-Net. This disparity is likely due to Mask R-CNN's more complex architecture and higher number of parameters, which require larger amounts of fine-tuning data to adapt effectively. With the small fine-tuning dataset available, Mask R-CNN struggles to update its deeper layers sufficiently, leading to poorer generalization and higher rates of false positives and false negatives. In contrast, Attention U-Net's simpler design and focus on segmentation through attention mechanisms allow it to adapt more efficiently with limited data, resulting in better performance after the same level of fine-tuning.

4.2. Inadequate Results on Small Object Segmentation

Detecting small shrubs using pre-trained models remains a significant challenge, even after extensive fine-tuning (Figure 7). The challenge exists in both Attention U-Net and Mask R-CNN models. For the Attention U-Net results shown in Figure 7b,c, although level two fine-tuning substantially improved the overall AP scores, the detection performance for small shrubs (canopy area < 3 m^2) remains relatively low, indicating persistent challenges in accurately segmenting small-scale objects. For the Mask R-CNN results shown in Figure 7d,e, fine-tuning led to moderate improvements in overall AP scores; however, the magnitude of improvement is notably smaller compared to that achieved by Attention U-Net, suggesting a more limited adaptation to the shrub detection task under the same fine-tuning conditions. Moreover, the model continues to exhibit relatively low performance in

detecting small shrubs (canopy area < 3 m^2), indicating limited sensitivity to small-scale objects despite fine-tuning.

The limited sensitivity to small shrubs (canopy area < 3 m^2) is likely attributed to the model's architectural and post-processing approach. Attention U-Net's encoder–decoder framework excels at preserving pixel-level details through high-resolution skip connections, which directly propagate shallow features (e.g., 256 × 256 resolution outputs from Layer L2) to the decoder pathway. This design maintains the fine-grained textures of small shrubs, a characteristic that is often lost due to the downsampling inherent in traditional networks [37]. Its simpler architecture and fewer parameters make it easier to train and more effective under limited data conditions. Furthermore, Attention U-Net incorporates self-attention gates (e.g., Layer L15) that dynamically recalibrate feature maps by modeling spatial dependencies between encoder and decoder activations [38], allowing the model to better capture small, subtle objects without needing a large number of examples. However, Attention U-Net tends to overgeneralize in low-density or early-season scenes, often missing smaller shrub crowns or blending them into the background. This issue is compounded when using the watershed post-processing method when intending to convert binary masks into shrub instances, which tends to overlook or eliminate small shrub detections due to their weaker edge definitions and limited spatial extent.

In contrast, Mask R-CNN with ResNet101-FPN backbone has a two-stage detection framework that imposes inherent limitations for detecting small shrubs. Although its Feature Pyramid Network (FPN) theoretically addresses multi-scale object recognition, the default practical resolution of deeper FPN layers (e.g., P4–P6) used in Ball et al.'s research [16] is insufficient for resolving the fine details of tiny shrubs. The Region Proposal Network (RPN) relies on anchor boxes designed for more common object sizes, which often fail to align with smaller shrubs due to their small footprint relative to the default anchor configurations (e.g., 32 × 32 anchors on P2 layers) [39]. Even with RoIAlign's sub-pixel accuracy [26], feature maps pooled from FPN layers for mask prediction lack the pixel-wise precision provided by Attention U-Net's skip connections, further compromising the detection of small shrubs. Moreover, Mask R-CNN often underperforms when fine-tuned on small training datasets due to its complexity and high parameter count. With limited data, the model is prone to overfitting and struggles to learn reliable region proposals, especially if the object shapes, sizes, and background textures in the small dataset are not diverse.

4.3. Low Model Sensitivity in the Early Growing Season

Both models exhibit limited sensitivity to shrubs in the early growing season (April to June; Figure 8). This is potentially due to shrubs often having sparse foliage in spring, limited canopy development, and maybe being mixed with background elements like soil, grasses, or early-season herbaceous plants. This lack of strong contrast in texture, color, and structure makes it difficult for deep learning models, especially those pre-trained for tree crown detection, to distinguish shrubs from the surrounding environment. As the season advances and the shrubs reach full maturity in late summer (July to September), their crowns become more defined, denser, and spectrally unique, allowing models to detect and delineate them with greater accuracy. The senesce of grass and forbs in fall also allows the still-active foliage on shrubs to stand out better across spectral ranges.

4.4. Potential Solutions to Enhance Segmentation Accuracy

4.4.1. Multispectral Integration

Although multispectral data was required in this study, two pre-trained models from previous research were developed using only RGB imagery. As a result, our fine-tuning

process was constrained to RGB inputs to maintain compatibility with the pre-trained weights and network architecture. To further improve model performance, one promising approach is to incorporate multispectral data, such as near-infrared and red-edge bands, which have been shown in previous research to enhance accuracy by approximately 5% [40]. This enhanced spectral information could improve the model's ability to detect subtle vegetation differences, leading to more accurate segmentation, especially in complex or heterogeneous landscapes.

In practice, integrating multispectral data requires modifying the input layers of deep learning models. For instance, in the Attention U-Net architecture, the input layer can be adapted to accept more than the standard three RGB channels by adjusting the number of input channels in the first convolutional layer from 3 to 5 to match the number of multispectral bands (RGB, NIR, and red-edge). Similarly, in Mask R-CNN, the first convolutional layer in its backbone should be modified from 3 to 5 to match the number of multispectral bands. This modification allows the model to process multispectral imagery, but would require training from scratch.

4.4.2. Model Architecture Adjustments

As mentioned in Section 4.2, the inadequate results in small shrub detection are fundamentally tied to their architectural designs. To address these issues, targeted architectural modifications were proposed. For Attention U-Net, integrating dilated convolutions in deeper encoder layers will preserve resolution without sacrificing context [41,42], while multi-scale pyramid attention enhances cross-stage feature fusion, amplifying faint object signals. Mask R-CNN benefits from denser, smaller anchors on P2/P3 FPN layers and FPN resolution boosting via a P1 layer to retain high-resolution features. Furthermore, training optimizations, such as higher input resolutions (2000×2000) and boundary-sensitive loss functions, could further boost precision [43].

Moreover, switching to models that perform well for small object segmentation, such as the high-resolution architecture HRNet [44], which maintains parallel multi-resolution branches throughout the network, and preserving full-resolution feature maps or hybrid CNN–Transformer models, like TransU-Net [45], which combines U-Net with Vision Transformers (ViTs) [46] in the encoder for global context modeling, would potentially improve the segmentation performance. However, these models may require training from scratch due to the unique characteristics of small object detection that require extensive data labeling and the creation of high-quality annotated datasets.

4.4.3. Synthetic Data Generation

Other than traditional data augmentation techniques, synthetic data generation offers a more advanced approach to enhancing small shrub segmentation, especially in data-scarce scenarios. By using generative models, such as Generative Adversarial Networks (GANs) [47], realistic synthetic imagery can be created that simulates various environmental conditions, including different lighting, weather, and seasonal changes. These synthetic images can be specifically designed to feature small shrubs with diverse shapes, sizes, and textures, which can help address the challenge of detecting small vegetation amidst complex backgrounds. Additionally, synthetic data allows for the creation of highly detailed annotations for training, ensuring that the model can learn to identify even subtle variations in shrub features. By augmenting real datasets with these synthetic examples, the model can improve its generalization capabilities, become more robust to various environmental factors, and ultimately achieve better performance in small shrub segmentation, especially in areas with scarce labeled data.

4.5. Applicability of Transfer Learning

This study highlights the potential of applying transfer learning in deep learning models to effectively monitor shrub cover using relatively high-resolution UAS imagery. The results demonstrate that, particularly during the mature stage of shrub growth, the models are capable of accurately detecting and delineating shrub crowns. While these models could enhance decisions on controlled burns and vegetation management in fragile ecosystems, significant challenges remain in translating these technologies into practical tools for land managers. Despite advancements in deep learning for ecological monitoring, a substantial gap exists between cutting-edge technology and its real-world application. Many land management agencies face barriers due to a lack of technical expertise, computational resources, and training to deploy these models. The complexity of model implementation, data preprocessing, and result interpretation further complicates widespread adoption. Future research in this area should emphasize the importance of collaboration between researchers and land managers to close this gap, with a focus on creating accessible, user-friendly tools that enable the practical application of these models in real-world conservation and land management efforts.

5. Conclusions

This study evaluated the transferability of pre-trained deep learning models across ecologically similar but structurally different areas for shrub segmentation using UAS-based remote sensing imagery. Our results highlight that fine-tuning is essential, especially when variations in shrub size, texture, and seasonal conditions are present. Among the two tested models, Attention U-Net and Mask R-CNN, we found that Attention U-Net is more effective when training resources are limited, offering better generalization with minimal annotated data. It is particularly suited for practitioners seeking rapid deployment with limited annotation efforts. Both models performed well in detecting mature shrubs during the late summer (July to September). However, segmentation accuracy dropped significantly for small shrubs with crown areas of less than 2 m^2 when imagery was captured at a 120 m flight elevation. These small objects often lacked distinct spatial features at this scale, leading to poor recall across both models. The challenge was even more pronounced for early-season (April to June) vegetation, where low contrast and undeveloped crowns made detection difficult. Future work should consider multi-scale feature fusion, denser anchor box strategies, and hybrid CNN–Transformer designs to improve small object segmentation. Additionally, integrating multispectral bands beyond RGB, such as near-infrared (NIR) and red-edge, could potentially improve model performance, especially for small or early-stage shrubs that are difficult to resolve using RGB alone. Also, flying at lower altitudes or using higher-resolution sensors could help overcome the current limitations in small shrub detection. These findings emphasize the need for improved model architectures and imaging strategies for accurately detecting and monitoring fine-scale vegetation structures, particularly small or early-stage shrubs, in complex and dynamic ecosystems.

Author Contributions: Conceptualization, D.B. and H.Z.; methodology, J.L.; software, J.L. validation, J.L.; investigation, J.L.; writing—original draft preparation, J.L.; writing—review and editing, D.B. and H.Z.; supervision, D.B. and H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: Oak Ridge Institute for Science and Education (ORISE) and USDA Agricultural Research Service SCINet Initiative, reference number USDA-ARS-SCINet-2023-0261.

Data Availability Statement: All datasets and code generated during this study are available upon request.

Acknowledgments: The authors thank Kevin Yemoto and Alex Olsen for collecting the UAS imagery and Rob Erskine for assistance in field operation.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
BN	Batch Normalize
CNNs	Convolutional Neural Networks
COCO	Common Objects in Context
IoU	Intersection over Union
Mask R-CNN	Mask Region-based Convolutional Neural Network
UASs	Unmanned Aircraft Systems
U-Net	U-shaped Network

Appendix A



Figure A1. The difference between (**a**) pre-trained model forest environments and (**b**) target shrubland environments.

Table A1. The augmentation methods used in this study; augmentation steps were executed with a stochastic application rate of 50%.

Type	Parameters
RandomBrightnessContrast	Brightness_limit (−0.1, 0.1), contrast_limit (−0.1, 0.1)
RandomToneCurve	Scale (0.1)
Downscale	Scale_range (0.5, 1)
AdvancedBlur	Blur_limit (3, 7)
Defocus	Radius (3, 10)
MotionBlur	Blur_limit (7) allow_shifted False
HorizontalFlip	N/A
RandomRotated90	N/A
RandomCrop	N/A

Table A2. Attention U-Net model architecture.

Stage No.	Layer No.	Layer Details	Output Shape
Input	L0	Input	(3, 256, 256)
Encoder:			
Stage 1	L1	Conv2d (3, 64, 3, 1, 1) + Relu	(64, 256, 256)
	L2	Conv2d (64, 64, 3, 1, 1) + Relu + BN	(64, 256, 256)
	L3	MaxPool2d (2)	(64, 128, 128)
Stage 2	L4	Conv2d (64, 128, 3, 1, 1) + Relu	(128, 128, 128)
	L5	Conv2d (128, 128, 3, 1, 1) + Relu + BN	(128, 128, 128)
	L6	MaxPool2d (2)	(128, 64, 64)
Stage 3	L7	Conv2d (128, 256, 3, 1, 1) + Relu	(256, 64, 64)
	L8	Conv2d (256, 256, 3, 1, 1) + Relu + BN	(256, 64, 64)
	L9	MaxPool2d (2)	(256, 32, 32)
Stage 4	L10	Conv2d (256,512,3,1,1) + Relu	(512, 32, 32)
	L11	Conv2d (512,512,3,1,1) + Relu + BN	(512, 32, 32)
	L12	MaxPool2d (2)	(512, 16, 16)
Stage 5	L13	Conv2d (512, 1024, 3, 1, 1) + Relu	(1024, 16, 16)
	L14	Conv2d (1024, 1024, 3, 1, 1) + Relu + BN	(1024, 16, 16)
Decoder:			
Bridge 4	L15	Self-attention block (L14, L11)	(1536, 32, 32)
	L16	Conv2d (1536, 512, 3, 1, 1) + Relu	(512, 32, 32)
	L17	Conv2d (512, 512, 3, 1, 1) + Relu + BN	(512, 32, 32)
Bridge 3	L18	Self-attention block (L17, L8)	(768, 64, 64)
	L19	Conv2d (768, 256, 3, 1, 1) + Relu	(256, 64, 64)
	L20	Conv2d (256, 256, 3, 1, 1) + Relu + BN	(256, 64, 64)
Bridge 2	L21	Self-attention block (L20, L5)	(384, 128, 128)
	L22	Conv2d (384, 128, 3, 1, 1) + Relu	(128, 128, 128)
	L23	Conv2d (128, 128, 3, 1, 1) + Relu + BN	(128,128, 128)
Bridge 1	L24	Self-attention block (L23, L2)	(192, 256, 256)
	L25	Conv2d (192, 64, 3, 1, 1) + Relu	(64, 256, 256)
	L26	Conv2d (64, 64, 3, 1, 1) + Relu + BN	(64, 256, 256)
	L27B	Conv2d (64, 1, 1, 1, 0)	(1, 256, 256)

Table A3. The self-attention block architecture used in Table A2, using L15 as an example.

Layer No.	Input Layer	Layer Details	Output Shape
A1	L14	Upsample (2)	(1024, 32, 32)
A2	L11	Conv2d (512, 256, 1, 1, 0)	(256, 32, 32)
A3	A1	Conv2d (1024, 256, 1, 1, 0)	(256, 32, 32)
A4	A2, A3	Add + Relu	(256, 32, 32)
A5	A4	Conv2d (256, 1, 1, 1, 0) + Sigmoid	(1, 32, 32)
A6	A5, L11	Multiply	(512, 32, 32)
A7	A6, A1	Concatenate	(1536, 32, 32)

References

1. Congalton, R.; Gu, J.; Yadav, K.; Thenkabail, P.; Ozdogan, M. Global Land Cover Mapping: A Review and Uncertainty Analysis. *Remote Sens.* **2014**, *6*, 12070–12093. [CrossRef]
2. Hansen, M.C.; Potapov, P.V.; Moore, R.; Hancher, M.; Turubanova, S.A.; Tyukavina, A.; Thau, D.; Stehman, S.V.; Goetz, S.J.; Loveland, T.R.; et al. High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science* **2013**, *342*, 850–853. [CrossRef] [PubMed]
3. Eldridge, D.J.; Bowker, M.A.; Maestre, F.T.; Roger, E.; Reynolds, J.F.; Whitford, W.G. Impacts of shrub encroachment on ecosystem structure and functioning: Towards a global synthesis: Synthesizing shrub encroachment effects. *Ecol. Lett.* **2011**, *14*, 709–722. [CrossRef] [PubMed]
4. Hellesen, T.; Matikainen, L. An Object-Based Approach for Mapping Shrub and Tree Cover on Grassland Habitats by Use of LiDAR and CIR Orthoimages. *Remote Sens.* **2013**, *5*, 558–583. [CrossRef]
5. Myint, S.W.; Gober, P.; Brazel, A.; Grossman-Clarke, S.; Weng, Q. Per-pixel vs. Object-based classification of urban land cover extraction using high spatial resolution imagery. *Remote Sens. Environ.* **2011**, *115*, 1145–1161. [CrossRef]
6. Zhang, L.; Du, B. Deep Learning for Remote Sensing Data: A Technical Tutorial on the State of the Art. *IEEE Geosci. Remote Sens. Mag.* **2016**, *4*, 22–40. [CrossRef]
7. Guirado, E.; Tabik, S.; Alcaraz-Segura, D.; Cabello, J.; Herrera, F. Deep-learning Versus OBIA for Scattered Shrub Detection with Google Earth Imagery: Ziziphus lotus as Case Study. *Remote Sens.* **2017**, *9*, 1220. [CrossRef]
8. James, K.; Bradshaw, K. Detecting plant species in the field with deep learning and drone technology. *Methods Ecol. Evol.* **2020**, *11*, 1509–1519. [CrossRef]
9. Khaldi, R.; Tabik, S.; Puertas-Ruiz, S.; de Giles, J.P.; Correa, J.A.H.; Zamora, R.; Segura, D.A. Individual mapping of large polymorphic shrubs in high mountains using satellite images and deep learning (Version 2). *arXiv* **2024**. [CrossRef]
10. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
11. Wei, W.; Cheng, Y.; He, J.; Zhu, X. A review of small object detection based on deep learning. *Neural Comput. Appl.* **2024**, *36*, 6283–6303. [CrossRef]
12. Panda, A.; Panigrahi, D.; Mitra, S.; Mittal, S.; Rahimi, S. Transfer Learning Applied to Computer Vision Problems: Survey on Current Progress, Limitations, and Opportunities (Version 1). *arXiv* **2024**. [CrossRef]
13. Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How transferable are features in deep neural networks? *arXiv* **2014**. [CrossRef]
14. Tuia, D.; Persello, C.; Bruzzone, L. Domain Adaptation for the Classification of Remote Sensing Data: An Overview of Recent Advances. *IEEE Geosci. Remote Sens. Mag.* **2016**, *4*, 41–57. [CrossRef]
15. Li, S.; Brandt, M.; Fensholt, R.; Kariryaa, A.; Igel, C.; Gieseke, F.; Nord-Larsen, T.; Oehmcke, S.; Carlsen, A.H.; Junttila, S.; et al. Deep learning enables image-based tree counting, crown segmentation, and height prediction at national scale. *PNAS Nexus* **2023**, *2*, pgad076. [CrossRef]
16. Ball, J.G.; Hickman, S.H.; Jackson, T.D.; Koay, X.J.; Hirst, J.; Jay, W.; Coomes, D.A. Accurate delineation of individual tree crowns in tropical forests from aerial RGB imagery using Mask R-CNN. *Remote Sens. Ecol. Conserv.* **2023**, *9*, 641–655. [CrossRef]
17. Kouw, W.M.; Loog, M. An introduction to domain adaptation and transfer learning (Version 2). *arXiv* **2018**. [CrossRef]
18. Singh, M.K.; Kumar, B. Fine Tuning the Pre-trained Convolutional Neural Network Models for Hyperspectral Image Classification Using Transfer Learning. In *Computer Vision and Robotics*; Shukla, P.K., Singh, K.P., Tripathi, A.K., Engelbrecht, A., Eds.; Springer Nature: Singapore, 2023; pp. 271–283. [CrossRef]

19. Lauenroth, W.K.; Burke, I.C. *Ecology of the Shortgrass Steppe: A Long-Term Perspective*; Oxford University Press: Oxford, UK, 2008.
20. Mahood, A.L.; Barnard, D.M.; Macdonald, J.A.; Green, T.R.; Erskine, R.H. Soil climate underpins year effects driving divergent outcomes in semi-arid cropland-to-grassland restoration. *Ecosphere* **2024**, *15*, e70042. [CrossRef]
21. Sun, C.; Shrivastava, A.; Singh, S.; Gupta, A. Revisiting unreasonable effectiveness of data in deep learning era. In Proceedings of the IEEE international conference on computer vision, Venice, Italy, 22–29 October 2017; pp. 843–852.
22. Lin, T.-Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C.L.; Dollár, P. Microsoft COCO: Common Objects in Context (Version 3). *arXiv* **2014**. [CrossRef]
23. Buslaev, A.; Parinov, A.; Khvedchenya, E.; Iglovikov, V.I.; Kalinin, A.A. Albumentations: Fast and flexible image augmentations. *arXiv* **2018**. [CrossRef]
24. Gishyan, K. Improving UAV Object Detection through Image Augmentation. *Math. Probl. Comput. Sci.* **2020**, *54*, 53–68. [CrossRef]
25. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization (Version 9). *arXiv* **2014**. [CrossRef]
26. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2961–2969. [CrossRef]
27. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
28. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125. [CrossRef]
29. Bottou, L. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT' 2010*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 177–186. [CrossRef]
30. Paszke, A.; Gross, S.; Malla, S.; Abbott, C.; Zhmoginov, A.; Chanan, G. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS) 32, Vancouver, BC, Canada, 8–14 December 2019.
31. Goyal, P.; Dollár, P.; Girshick, R.; Noordhuis, P.; Wesolowski, L.; Kyrola, A.; He, K. Accurate, large minibatch SGD: Training ImageNet in 1 hour. *arXiv* **2008**, arXiv:1706.02677.
32. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The PASCAL Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis. (IJCV)* **2010**, *88*, 303–338. [CrossRef]
33. Efron, B. Bootstrap Methods: Another Look at the Jackknife. In *Breakthroughs in Statistics: Methodology and Distribution*; Kotz, S., Johnson, N.L., Eds.; Springer: Berlin/Heidelberg, Germany, 1992; pp. 569–593. [CrossRef]
34. Zhang, S.; Zou, X.; Li, K.; Lang, C.; Wang, S.; Tao, P.; Cao, T. Knowledge Transfer and Domain Adaptation for Fine-Grained Remote Sensing Image Segmentation (Version 3). *arXiv* **2024**. [CrossRef]
35. Ma, X.; Zhang, X.; Pun, M.-O.; Huang, B. MANet: Fine-Tuning Segment Anything Model for Multimodal Remote Sensing Semantic Segmentation (Version 1). *arXiv* **2024**. [CrossRef]
36. Ayhan, B.; Kwan, C. Tree, Shrub, and Grass Classification Using Only RGB Images. *Remote Sens.* **2020**, *12*, 1333. [CrossRef]
37. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer International Publishing: Cham, Switzerland, 2015; Volume 9351, pp. 234–241. [CrossRef]
38. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need (Version 7). *arXiv* **2017**. [CrossRef]
39. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
40. Ayhan, B.; Kwan, C.; Budavari, B.; Kwan, L.; Lu, Y.; Perez, D.; Li, J.; Skarlatos, D.; Vlachos, M. Vegetation Detection Using Deep Learning and Conventional Methods. *Remote Sens.* **2020**, *12*, 2502. [CrossRef]
41. Quan, W.; Zhao, W.; Wang, W.; Xie, H.; Lee Wang, F.; Wei, M. Lost in UNet: Improving Infrared Small Target Detection by Underappreciated Local Features. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 1–15. [CrossRef]
42. Wang, S.; Singh, V.K.; Cheah, E.; Wang, X.; Li, Q.; Chou, S.-H.; Lehman, C.D.; Kumar, V.; Samir, A.E. Stacked dilated convolutions and asymmetric architecture for U-Net-based medical image segmentation. *Comput. Biol. Med.* **2022**, *148*, 105891. [CrossRef]
43. Du, J.; Guan, K.; Liu, P.; Li, Y.; Wang, T. Boundary-Sensitive Loss Function With Location Constraint for Hard Region Segmentation. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 992–1003. [CrossRef] [PubMed]
44. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep High-Resolution Representation Learning for Visual Recognition (Version 2). *arXiv* **2019**. [CrossRef]
45. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation (Version 1). *arXiv* **2021**. [CrossRef]

46. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale (Version 2). *arXiv* **2020**. [CrossRef]
47. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks (Version 1). *arXiv* **2014**. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.