drones

MDPI

Article

Large-Scale Date Palm Tree Segmentation from Multiscale UAV-Based and Aerial Images Using Deep Vision Transformers

Mohamed Barakat A. Gibril [1,2], Helmi Zulhaidi Mohd Shafri [1,*], Rami Al-Ruzouq [2,3], Abdallah Shanableh [2,3], Faten Nahas [4] and Saeed Al Mansoori [5]

1 Department of Civil Engineering and Geospatial Information Science Research Centre (GISRC), Faculty of Engineering, Universiti Putra Malaysia (UPM), Serdang 43400, Selangor, Malaysia
2 GIS and Remote Sensing Center, Research Institute of Sciences and Engineering, University of Sharjah, Sharjah 27272, United Arab Emirates
3 Department of Civil and Environmental Engineering, Faculty of Engineering, University of Sharjah, Sharjah 27272, United Arab Emirates
4 Geography Department, College of Arts, King Saud University, Riyadh 1145, Saudi Arabia
5 Remote Sensing Department, Mohammed Bin Rashed Space Center (MBRSC), Dubai 211833, United Arab Emirates
* Correspondence: helmi@upm.edu.my; Tel.: +60-3-97696459

Abstract: The reliable and efficient large-scale mapping of date palm trees from remotely sensed data is crucial for developing palm tree inventories, continuous monitoring, vulnerability assessments, environmental control, and long-term management. Given the increasing availability of UAV images with limited spectral information, the high intra-class variance of date palm trees, the variations in the spatial resolutions of the data, and the differences in image contexts and backgrounds, accurate mapping of date palm trees from very-high spatial resolution (VHSR) images can be challenging. This study aimed to investigate the reliability and the efficiency of various deep vision transformers in extracting date palm trees from multiscale and multisource VHSR images. Numerous vision transformers, including the Segformer, the Segmenter, the UperNet-Swin transformer, and the dense prediction transformer, with various levels of model complexity, were evaluated. The models were developed and evaluated using a set of comprehensive UAV-based and aerial images. The generalizability and the transferability of the deep vision transformers were evaluated and compared with various convolutional neural network-based (CNN) semantic segmentation models (including DeepLabV3+, PSPNet, FCN-ResNet-50, and DANet). The results of the examined deep vision transformers were generally comparable to several CNN-based models. The investigated deep vision transformers achieved satisfactory results in mapping date palm trees from the UAV images, with an mIoU ranging from 85% to 86.3% and an mF-score ranging from 91.62% to 92.44%. Among the evaluated models, the Segformer generated the highest segmentation results on the UAV-based and the multiscale testing datasets. The Segformer model, followed by the UperNet-Swin transformer, outperformed all of the evaluated CNN-based models in the multiscale testing dataset and in the additional unseen UAV testing dataset. In addition to delivering remarkable results in mapping date palm trees from versatile VHSR images, the Segformer model was among those with a small number of parameters and relatively low computing costs. Collectively, deep vision transformers could be used efficiently in developing and updating inventories of date palms and other tree species.

Keywords: vision transformer; semantic segmentation; tree crown delineation; Segformer; Swin transformer; Segmenter; dense prediction transformer; CNN

Citation: Gibril, M.B.A.; Shafri, H.Z.M.; Al-Ruzouq, R.; Shanableh, A.; Nahas, F.; Al Mansoori, S. Large-Scale Date Palm Tree Segmentation from Multiscale UAV-Based and Aerial Images Using Deep Vision Transformers. *Drones* **2023**, *7*, 93. https://doi.org/10.3390/drones7020093

Academic Editors: Eben Broadbent and David R. Green

Received: 20 December 2022
Revised: 17 January 2023
Accepted: 26 January 2023
Published: 29 January 2023

1. Introduction

The date palm (*Phoenix dactylifera* L.), which is an arborescent monocotyledonous tree species with unique anatomical features (a single trunk, palm leaves, and fronds), is one of the most crucial fruit trees in arid and semi-arid regions. The crown of a date palm tree

is densely covered with long, pinnate leaves, some of which may reach a length of four meters [1], depending on the tree's cultivar, age, and climate conditions. Date palm trees normally grow to a height of between 15 m and 25 m [2]. It was originally grown in the Middle East and North Africa and has successfully been domesticated and spread to other areas with a suitable climate and sufficient water resources [3]. The annual production of the date palm exceeds 9 million tons, cultivated from over 1.38 million hectares [4]. Given the socioeconomic importance of date palm trees, it is essential to monitor, preserve, and ensure the precise management of date production.

Because date palm trees are distributed across vast agricultural and urban areas, very-high spatial resolution (VHSR) remotely sensed data can be utilized to map, detect, and count date palm trees. Over the past few years, the increasing availability of UAV-based images has considerably contributed to the development and the success of artificial intelligence models that accurately delineate individual tree crowns [5–10], count the number of trees [11–14], and assess the health of the trees [15,16]. Deep learning models, especially convolutional neural networks (CNNs), have been proven to have excellent performance in extracting tree crowns and mapping tree species from UAV data through patch-based classification [17–20], object detection [21–24], semantic segmentation [25–27], and instance segmentation [28–31].

Semantic segmentation is one of the key pillars in computer vision; it assigns a label to each image pixel. Numerous fully convolutional networks (FCNs), such as SegNet [32], U-Net [33], DeepLab V3+ [34], PSPNet [35], and others, have been adopted in forestry and precision agriculture domains to delineate the tree crowns from various sources of remotely sensed data. For instance, Gibril et al. [36] proposed a deep U-Net based on the residual learning framework (ResNet) for the large-scale mapping of date palm trees from UAV images. The proposed approach surpassed several state-of-the-art FCNs and achieved an F-score and mean IoU of 91% and 85%, respectively. Anagnostis et al. (2021) [37] utilized a U-Net model to detect and localize orchard trees from aerial images with various conditions (i.e., different seasons, different tree ages, and different levels of weed coverage) and achieved an accuracy of 91%, 90%, and 87% for training, validation, and testing, respectively. Ferreira et al. (2021) [38] trained DeepLabv3+ architecture with three backbones (ResNet-18, ResNet-50, and MobileNetV2) to map Brazilian nut trees from WorldView-3 satellite imagery. The authors reported an F-score ranging from 60.9% to 71.81%. Generally, the commonly used semantic segmentation networks for tree crown mapping are U-Net [39–46] and DeepLabV3+ [47–50], which have diverse backbone networks.

Given their remarkable performance in natural language processing, transformers [51] have been adopted in computer vision and have achieved outstanding performance. Vision transformers (ViT) use a pure transformer architecture, which is an effective backbone network that learns global features through a self-attention mechanism, as a feature extractor in image recognition [52], object detection [53,54], and semantic segmentation tasks [55]. The input images are subdivided into small square patches (which are treated in the same way as tokens). These small patches are then flattened into vectors and are discretely embedded using linear projection. The positional embedding of these small patches is incorporated in order to preserve their positional information. Then, the sequence is fed into numerous multi-head attention layers, which generate the final representation. Transformers have larger receptive fields than CNN-based models and a higher representation power, and they can attain global contextual information through self-attention mechanisms [56]. Recently, several remote sensing studies have explored the potential of using several vision transformers in image classification [57,58], object detection [59,60], semantic segmentation [61–66], and instance segmentation [67].

In the context of individual tree crown mapping and detection, the Swin transformer (which is a hierarchal vision transformer) was adopted as the backbone of a variety of deep learning architectures. Gibril et al. [68] evaluated the performance and the generalizability of various instance segmentation algorithms with varying backbones in detecting and

mapping individual date palm trees from multiscale UAV images. The efficiency of various backbones, including residual learning networks (ResNets) and the Swin transformer, was scrutinized. The mask-region convolutional neural network that was based on the Swin transformer backbone demonstrated outstanding performance, and it outperformed those with different ResNet architectures and depths. The superiority of the Swin transformer was attributed to its ability to attain global contextual information through the gradual increase in its receptive field. Lan et al. [69] proposed a lightweight object detection model that was based on the YOLOX detector and the tiny version of the Swin transformer as a backbone network for the real-time UAV patrol orchard mission. The YOLOX network that was based on the Swin transformer backbone showed an improvement in its accuracy (by 1.9%) and its size over the original YOLOX network. Alshammari and Shahin [70] adopted a SwinTU-net, which is an encoder–decoder semantic segmentation model that is based on a U-Net design and a Swin transformer backbone, for segmenting olive trees from UAV images. The authors' experimental results showed that the performance of the SwinTU-net model surpassed similar studies in identifying olive trees.

The vast majority of previous studies on date palm have focused on developing and evaluating CNN-based models using a single data source and limited study areas. The large-scale mapping of date palm trees from multisource VHSR images can be challenging because different imaging systems acquire remotely sensed data with varying spatial and radiometric resolutions. Moreover, the limited spectral information of VHSR images, the high intraclass variance of date palm trees (i.e., the cultivar, the age, the crown size, and the height), and the complexity of the image backgrounds can contribute to decreasing the generalizability of the deep learning models. Given their ability to capture robust local–global feature representations, which can boost the accuracy of palm tree mapping, the adoption of transformer-based models is proposed in this study to address the aforementioned challenges. Although a broad spectrum of CNNs demonstrates excellent performance in delineating the crowns of different tree species from various sources of remotely sensed data, the effectiveness and the efficiency of varying vision transformers in extracting tree species from multiscale and multisource VHSR images have not been fully explored.

The present study mainly aims to provide an end-to-end, efficient, and transferable deep learning approach for the large-scale mapping of date palm trees from multiscale UAV-based aerial images. The contributions of this study can be summarized as follows: (1) exploring and evaluating the performance of numerous state-of-the-art transformer-based DL models (i.e., the Segformer, the Segmenter, the UperNet-Swin, and the dense prediction transformer) with various levels of model complexities in the delineation of date palm trees; (2) assessing the generalizability of the deep transformer-based models on multiple testing data with varying spatial resolutions; (3) conducting an in-depth analysis on the transferability of the evaluated vision transformers and various CNN-based models (i.e., DeepLabV3+, PSPNet, FCN-ResNet-50, and DANet) to other VHSR images that were acquired from different geographical regions.

2. Materials

2.1. Study Area

The study area, which is shown in Figure 1, is positioned in the eastern region of Ajman Emirate, United Arab Emirates (UAE). The study site covers 30 km^2 and is characterized by diverse agricultural farms and residential areas. The date palm trees in the study area are planted in agricultural and urban landscapes and vary in density, cultivar, age, crown size, and height. The climatic conditions in the area range from arid to hyper-arid, with average temperatures between 18 °C and 34 °C, and a maximum temperature of over 40 °C in the summer season.

Figure 1. Study area: (**a**) UAE map; (**b**) location of the study area; (**c**–**e**) an example of multisource images.

2.2. Data Acquisition

This study utilized multisource and multiscale RGB images for training and evaluating various deep vision transformer models for the large-scale segmentation of date palm trees. The datasets consist of large-scale UAV orthomosaic images and two very-high-resolution aerial photographs.

The UAV dataset was acquired in the winter of 2019 between 9:00 a.m. and 1:00 p.m., using a SenseFly eBee-Plus UAV system equipped with a S.O.D.A. (sensor optimized for drone applications) digital camera. The camera captures visible-color images (RGB) with a resolution of 1.2 MP and an image size of 5472 × 3648 pixels (3:2). It can capture images in a variety of lighting conditions (e.g., sunny, cloudy, and shady) due to its automatic white balance feature. To cover the entire study site, the area was divided into small blocks (i.e., small flight missions) because of the limited battery time (a maximum endurance of 59 min) and range of the radio communication links of the UAV system.

The flight missions were planned, controlled, and visualized using SenseFly eMotion software. The flight missions were carried out at an average height of 100 m above elevation data, with frontal and lateral image overlaps of 85% and 80%, respectively. A survey-grade GNSS receiver was also set up as a reference station to concurrently record the data during the flights in a static mode. The very-high-resolution aerial images used in this study were acquired in the summer of 2017 and 2015, with a spatial resolution of 20 cm and 15 cm, respectively.

3. Methodology

The framework of this study, which is shown in Figure 2, consists of the following five main stages: (1) data preprocessing, (2) data preparation and split (Section 3.2), (2) model experiments (Section 3.5), (3) generalizability evaluation (Section 3.6), (4) transferability analysis (Section 3.7), and (5) application to large-scale images (Section 3.8).

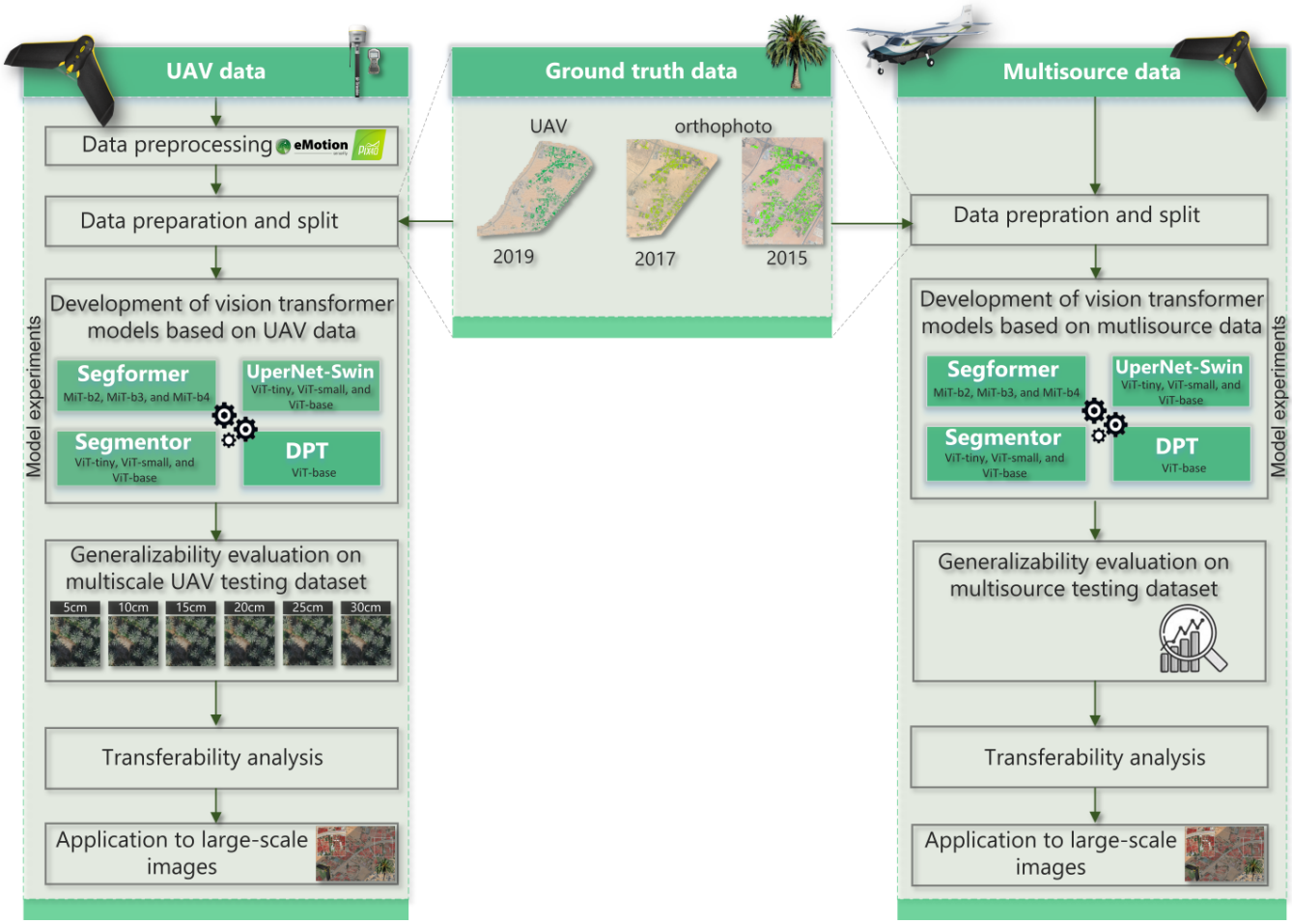

Figure 2. Methodological framework of this study.

3.1. Data Preprocessing

In this study, the UAV GNSS data (UAV flight log files) and the ground GNSS RINEX (receiver-independent exchange format) data were processed in eMotion software to rectify the positions of the UAV where the images were obtained during the flights and obtain geotagged images. After completing the necessary corrections, the geotagged images were imported into Pix4D mapper software (v.4.4.10; Prilly, Switzerland) to generate one orthomosaic RGB image with an average ground sampling distance (GSD) of 5 cm.

3.2. Data Preparation and Split

Developing a powerful deep-learning model for the segmentation of date palm trees requires high-quality and adequate ground-truth data. For the purpose of developing a generic and transferable deep learning model, comprehensive mutlidate vector layers were manually prepared by encircling the pixels of the crowns of the date palm trees (areas occupied by palm leaves) from the multisource RGB images based on a visual assessment. Another analyst meticulously checked and improved the vector layers to ensure that the date palm tree crowns were delineated properly. Overall, the prepared vector data consists of 50,000 date palm trees (prepared from the UAV data with a GSD of 5 cm) and 65,740 and 52,335 date palm trees (prepared from two aerial images with a GSD of 15 cm and 20 cm,

respectively). The revised multiscale vector layers of the date palm trees were converted into raster layers (binary masks).

In this study, three distinct regions, which were covered with varying types and sizes of date palm trees (Figure 1), were dedicated to training, validating, and testing deep vision transformer models. The site dedicated to training accounts for about 75% of the areas covered with palm trees, while the remaining 25% were kept for validating and testing the performance of the segmentation models. The remotely sensed data of the three regions and the corresponding mask were divided into equal-sized (512 × 512) image–mask pairs. Given the differences in the spatial resolutions of the various datasets, the number of generated image tiles varied from one dataset to another. A total of 21,045 UAV image tiles and their corresponding masks were obtained from the training region. A set of 3784 UAV image–mask pairs were obtained from the validation area, while a total of 3491 image–mask pairs were obtained from the testing region. Table 1 lists the number of generated images from each dataset using the training, validation, and testing regions. Figure 3 shows the samples of the generated multisource image tiles and their corresponding masks.

Table 1. A summary of the multiscale image tiles generated from the training, validation, and testing regions.

	UAV Data	Orthophoto	Orthophoto	Total
Acquisition	2019	2017	2015	-
Spatial resolution	5 cm	20 cm	15 cm	-
Training samples	21,045	5687	16,531	43,263
Validation samples	3784	221	682	4687
Testing samples	3491	218	657	4366

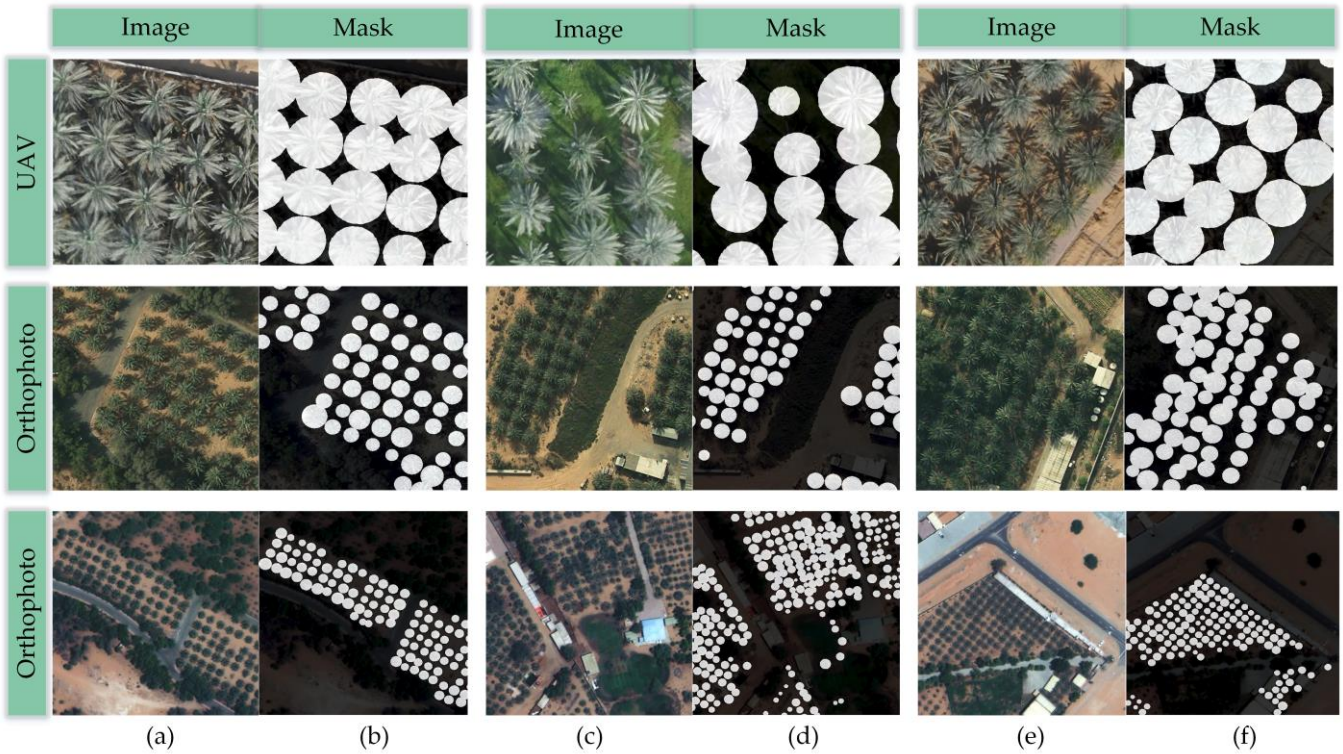

Figure 3. Illustrations of image–mask pairs: (**a**,**c**,**e**) multisource image tiles and (**b**,**d**,**f**) their corresponding masks.

3.3. Semantic Segmentation Models

This research investigated the performance of various deep vision transformers on multiscale UAV and multisource datasets. First, the deep learning models were trained and evaluated using UAV data (with 5 cm GSD). The trained models were used to evaluate the performance of the models in mapping the date palm trees from multiscale UAV data with various GSD (i.e., 10, 15, 20, 25, and 30 cm). Second, the performance of the deep learning models trained on multiscale RGB images (UAV and two aerial images) was also evaluated. The following subsections provide a brief description of the utilized deep vision transformers.

3.3.1. Segformer

The Segformer [71], which is a powerful and efficient transformer-based semantic segmentation framework, is based on an encoder–decoder architecture (Figure 4). The encoder includes a set of four transformer blocks for multiscale feature learning. The transformer blocks comprise the following three modules: efficient self-attention, a mixed feed-forward network (Mix-FFN), and overlap patch merging blocks. The input images (i.e., h, w, and c) are divided into small patches (i.e., 4×4), and the patches are passed to a hierarchical transformer encoder to extract multi-level features in order to boost the performance of semantic segmentation. The efficient self-attention module employs a sequence reduction technique [72] to lessen the computational cost. Unlike traditional ViT, which utilizes fixed resolution position encodings (PEs) to incorporate positional information, the Segformer uses convolutional layers in the feed-forward network (FFN) for data-driven positional encoding. The Mix-FFN provides positional information by considering the effect of zero padding to leak location information [69] by using 3×3 convolutions in the FFN [73]. The overlap patch merging module generates feature maps of the same process size without overlapping, resulting in hierarchical feature representation representing local and global information. The multiscale features, which include low-resolution coarse features, and high-resolution fine features, are aggregated at resolutions of $1/4$, $1/8$, $1/16$, and $1/32$ of the original images. An example of multiscale feature maps generated by the Segformer is shown in Figure 5. The complexity of the Segformer depends on the selection of the depth of the mix transformer (MiT) encoder, which varies from B0 to B5 [71]. This study investigated the performance of MiT B2, B3, and B4 encoders.

The decoder encompasses four main steps and employs lightweight multi-layer perceptrons (MLPs) that combine local and global attention to fuse multiscale feature outputs from different layers and predict the final segmentation results. First, the extracted multi-level features are fed to the MLP layer to unify the channel dimension. Second, the features are upsampled to $1/4$ and concatenated. Third, the concatenated feature maps are fused by an MLP layer, and the fused features are ultimately passed to another MLP layer to predict the segmentation mask. The decoder design is unique and produces a robust representation without computationally intensive modules.

3.3.2. UperNet-Swin Transformer

The design of the unified perceptual parsing network (UperNet) [74] is an encoder–decoder semantic segmentation network based on a feature pyramid network (FPN) [75] and a pyramid pooling module (PPM) from PSPNet [35]. The FPN constructs bottom-up (downsampling feature maps) pathways, top-down (upsampling feature maps) pathways, and lateral connections. While the bottom-up pathway extracts multiscale feature maps (depending on the selected backbone network), the top-down pathway increases the spatial dimensions of the feature maps by a factor of two. The PPM is applied to the final layer of the bottom-up network before it is fed into the top-down pathway via lateral connections. Through the lateral connections, the top-down and bottom-up feature maps are integrated to create low- (semantically strong features) and high-resolution

(semantically weak features) feature maps. Additional information on the FPN and PPM can be found in [35,75].

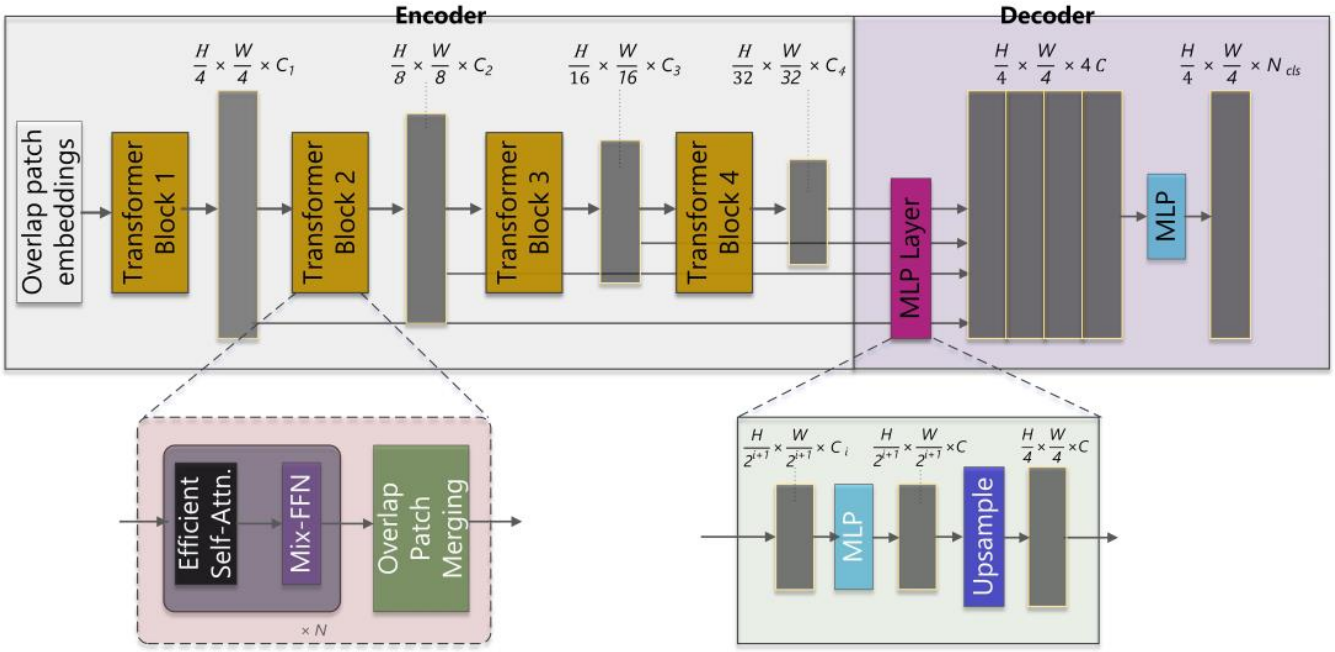

Figure 4. Framework of the Segformer.

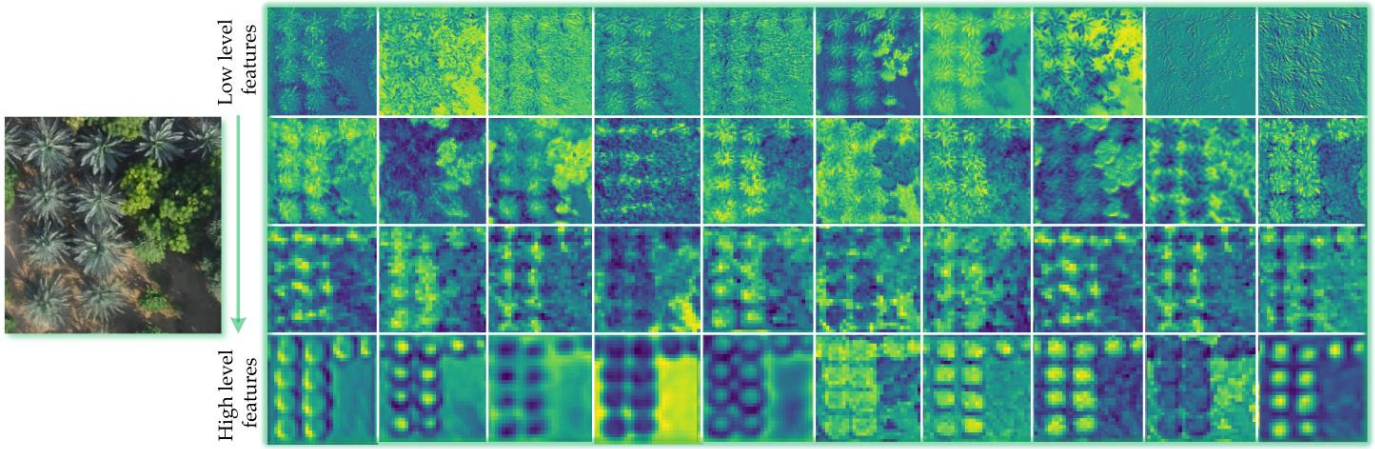

Figure 5. Examples of the multiscale feature maps extracted by the Segformer.

In this study, the Swin transformer was used as the encoder of UperNet. The Swin transformer [76] is an effective multistage hierarchical vision transformer architecture whose representation is obtained via the shifted window scheme, which is used to extract multi-level features and model the long-range dependencies in the data. The shifted windowing technique improves efficiency by confining self-attention computation to non-overlapping local windows while enabling cross-window connections. The architecture of the Swin transformer encompasses four stages, including several operations to extract multiscale feature maps, patch partition, patch merging, linear embedding, and Swin transformer blocks, as shown in Figure 6.

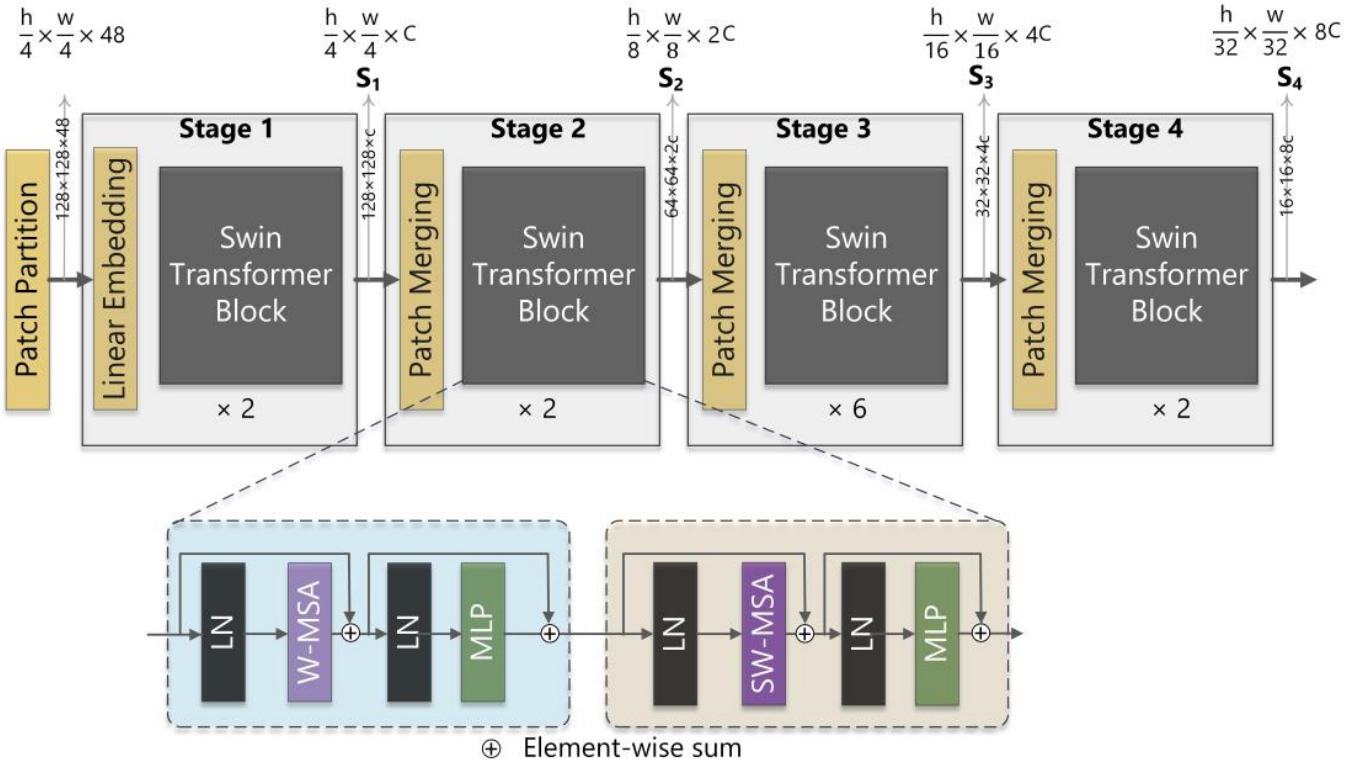

Figure 6. The encoder is based on Swin Transformer (Swin-tiny version).

Similar to ViT, the input image is partitioned into non-overlapping patches (i.e., 4×4), and each patch is regarded as a token. The feature of each token is set as the concatenation of raw multichannel pixel values. For instance, the feature dimension of a 4×4 patch is $4 \times 4 \times 3 = 48$ pixels. The raw-valued feature is projected into an arbitrary dimension (labeled as C) by the linear embedding layer. For the different variants of the Swin transformer, tiny, small, and base models, C is 96, 96, and 128, respectively. The patch tokens (feature vectors) are then passed through a set of transformer blocks for feature representation learning. A Swin transformer block comprises window multi-head self-attention (W-MSA), shifted window multi-head self-attention, and multi-layer perceptron (MLP). A LayerNorm (LN) layer is employed before each MSA module and MLP. In the interim, a residual connection is implemented after every module. After each stage, a patch merging layer is used to reduce the number of tokens and produce a hierarchical representation. For the Swin-tiny, Swin-small, and Swin-base models, the number of layers in stages 1, 2, 3, and 4 are {2, 2, 6, 2}, {2, 2, 18, 2}, and {2, 2, 18, 2}, respectively. This study investigated the performance of Swin-tiny, Swin-small, and Swin-base encoders.

3.3.3. Segmenter

The Segmenter [77], which is a transformer-based encoder–decoder architecture, maps a series of patch embeddings to pixel-level class annotations. The encoder is constructed using the vision transformer ViT [52] and considers all of the ”tiny”, ”small”, ”base,” and ”large” models. The input images are divided into a sequence of patches. Each patch is flattened into a 1D vector and linearly projected to a patch embedding to generate a sequence of patch embeddings. Learnable position embeddings are incorporated into the sequence of patches in order to retain the positional information. The transformer encoder consists of a set of transformer layers. Each layer encompasses a multi-headed self-attention (MSA) block, a two-layer point-wise MLP block with layer norm (LN) applied before each block, and residual connections that are added after each block. The decoder (mask transformer) maps the encoder’s patch-level encodings to the patch-level class

scores. These patch-level class scores are then upsampled to pixel-level scores via bilinear interpolation. Additional information on the Segmenter can be found in [77].

3.3.4. Dense Prediction Transformer

The dense prediction transformer (DPT) [78], which is an encoder–decoder architecture, leverages vision transformers [52] as a backbone for dense prediction tasks and as a convolutional decoder. The input images are subdivided into image patches, and then the patches are flattened into vectors and individually embedded using a linear projection. The image embeddings are coupled with a learnable position embedding to incorporate this information into the representation. The tokens are transformed by employing consecutive blocks of multi-headed self-attention (MHSA) [51] that link the tokens to each other and transform the representation. Given that the tokens have a one-to-one correspondence with image patches, the transformer maintains the number of tokens throughout all computations; thus, the ViT encoder preserves the spatial resolution of the initial embedding across all transformer stages. A three-stage reassemble operation is used to restore image-like representations from the output tokens of the arbitrary layers of the transformer encoder [78]. Tokens from the different stages of the vision transformer are assembled into image-like representations at various resolutions and are merged progressively into full-resolution predictions using a convolutional decoder.

3.4. Evaluation Metrics

This study used various evaluating metrics to quantify and analyze the effectiveness of different deep vision transformers in segmenting date palm trees from multisource remotely sensed data. The evaluating metrics include precision, recall, mean accuracy (mAcc), mean intersection-over-union (mIOU), and mean F-score (mF-score). These metrics have been widely employed for segmentation evaluation, which computes the degree of agreement between the manually annotated masks and the semantically segmented pixels by transformers. This can be expressed mathematically by Equations (1)–(6).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \tag{3}$$

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

$$\text{mIoU} = \frac{1}{2}(\text{IoU}_{\text{backgrounds}} + \text{IoU}_{\text{date palm tree}}) \tag{5}$$

$$\text{F} - \text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

where TP, TN, FP, and FN denote the true positive, true negative, false positive, and false negative, respectively.

3.5. Experimental Setup

The experiments were carried out using Pytorch [79] and MMsegmentation frameworks [80]. The deep learning models were trained on a Linux cluster with an Intel Xeon processor, eight NVIDIA Tesla K80 graphics processing units, 512 GB of RAM, and 100 TB of storage. The encoder of all of the models was initialized using ImageNet pretrained weights. The optimizer and loss functions were the AdamW (Adam with decoupled weight decay) and the cross-entropy loss, respectively. The initial learning rate, coefficients of momentum, and weight decay were assigned as 0.001, 0.9, and 0.0005, respec-

tively. The batch size was set to utilize the eight GPUs fully. The models were trained for 50,000 iterations and evaluated every 2000 iterations during the training using the validation dataset, and the best models were used in further analysis.

3.6. Generalization Capability Analysis

The spatial resolution of the VHSR data is one of the key factors that affect the appearance of date palm trees in the remotely sensed data, thereby limiting the generalizability of the deep learning models. In this study, the evaluated deep vision transformers were first trained on UAVs with a ground space distance of 5 cm. The generalizability of the trained models in mapping date palm trees from multiscale UAV images was assessed on different testing UAV datasets with various s (e.g., 10, 15, 20, 25, and 30 cm).

3.7. Transferability Analysis

The models were trained on the UAV dataset alone and were also trained on the combined multisource data. To elucidate the transferability of the evaluated deep vision transformers, the performance of the models was scrutinized on an independent public dataset [81]. The independent UAV dataset was acquired in Al Ain City with a spatial resolution of 10 cm. A total of 1078 image tiles (containing 9883 date palm trees) and their corresponding annotation of the date palm tree structure in Pascal VOC format (bounding boxes of the palm trees in XML format) [82] were selected. In this study, the crown of the date palm trees in these UAV data was manually demarcated using the LabelMe software (an open-source python annotation tool). The annotations were then converted into binary masks using *labelme2voc.py* [82] and utilized to assess the transferability of the various semantic segmentation models.

3.8. Model Application to Large-Scale Images

Developing deep learning models requires the VHSR images to be divided into equal image tiles (i.e., 512 $\times$ 512) because of the enormous size of the remotely sensed data and the formidable amount of computation required (Section 3.2). Following the training and evaluation stages of the different segmentation models, the process of employing the developed segmentation model to map the date palm trees from large-scale images involved the division of the large image into small, geotagged image tiles by using the GDAL library (i.e., GDAL retile). The generated tiles were then passed to the segmentation models, and the results were saved as binary geotagged images. Then, a virtual dataset (VRT, i.e., a mosaic form of the lists of the geotagged results) was built, and the final mosaic of the result was generated (i.e., GDAL translate module). The results were converted to vector data by applying watershed or multiresolution segmentation, followed by a threshold, and converted again to a vector layer (Shapefile or Geojason). Alternatively, the *measure.find_contours* module could be applied in the scikit-image library to segregate and extract the coordinates of the boundary of the date palm trees, and the coordinates could be saved to Geojason/Shapefile by using the Fiona or Shapely libraries.

4. Results

4.1. Evaluation of Vision Transformers Based on UAV Data

4.1.1. Results of Date Palm Tree Segmentation

This study investigated the potential of different vision transformers with varying complexities, including the Segformer, the Segmenter, the Swin transformer, and the DPT, for large-scale date palm tree mapping. The segmentation models were trained and evaluated based on a comprehensive UAV dataset. The semantic segmentation metrics, which includes the mIoU, the F-measure, and the mAcc of the different transformer-based models on the testing dataset, are shown in Figure 7. The evaluated models successfully delineated the date palm trees with the mIoU, the mF-score, and the mAcc ranging from 85–86.3%, 91.6–92.4%, and 91.3–92.5%, respectively. Figure 8 depicts a set of examples that were selected from the testing dataset, their corresponding ground-truth label, and

the results of the Segmenter with ViT-small, the DPT with ViT-base, the UperNet with Swin-tiny, and the Segformer with MiT-b2.

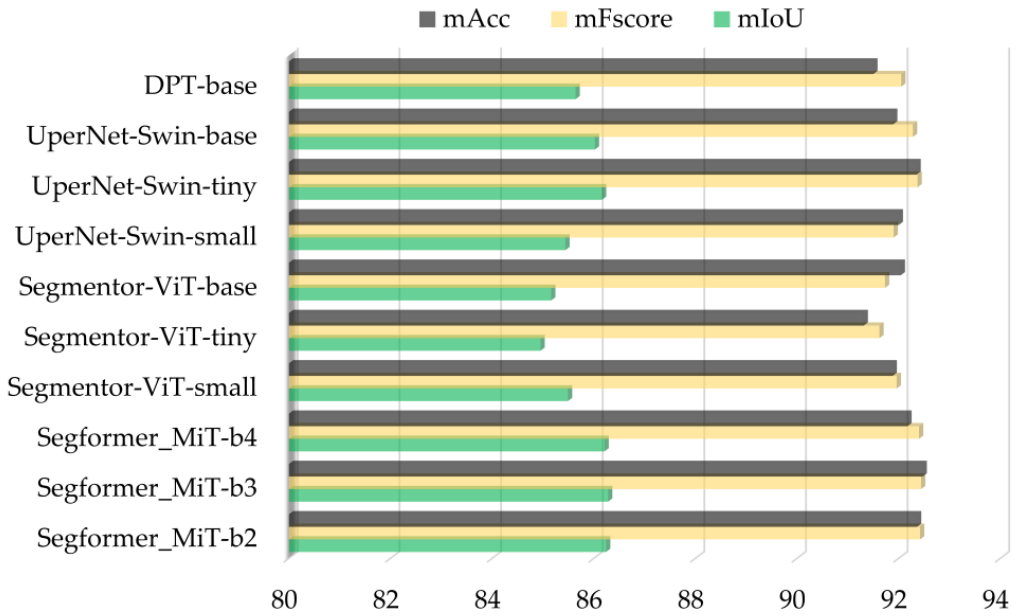

Figure 7. Segmentation accuracy metrics obtained from UAV testing data.

4.1.2. Results of Generalizability Evaluation

The acquisition of UAV data at large scale may necessitate flying at different altitudes in different regions and would depend on the granted flight permissions; thus, the spatial resolution of the data varies accordingly. The accuracy of the semantic segmentation models for date palm tree mapping is expected to drop with the decrease in the spatial resolution of the data. Thus, the evaluated vision transformers were trained on the UAV dataset with a ground space distance of 5 cm. The models were evaluated by testing the UAV datasets with different GSDs (i.e., 10, 15, 20, 25, and 30 cm) in order to assess the generalizability of the segmentation models on multiscale UAV data. Figure 9 lists the results of the generalizability evaluation. The evaluated vision transformers, except for the Segmenter, maintained an almost similar range of mIoU when they were applied to a GSD of 10 cm. When the models were applied to a testing dataset with a GSD of 15 cm, the Segformer-MiT-b4 and the UperNet-Swin-small outperformed the rest of the models and scored an mIoU of 80.74% and 80.09%, respectively. The mIoU was reduced by 11% to 23.5% when the spatial resolution of the data was reduced by 1/4 (20 cm). The Segformer and the UperNet-Swin demonstrated excellent generalization capacities for mapping date palm trees from multiscale UAV data (with a GSD ranging from 5 to 20 cm).

4.2. Results of Vision Transformers Developed on Multisource Data

The deep vision transformers were trained and evaluated on a comprehensive multi-date and multisource remotely sensed data (including a UAV-data (GSD of 5 cm) and two orthophotos (GSD of 15 and 20 cm)) in order to develop a generic semantic segmentation model for mapping date palm trees from different sources of VHSR images that enables creating and updating palm tree inventories. Figure 10 displays the mIoU, the mFscore, and the mAcc that were generated from the combined testing dataset (Table 1). Among the evaluated vision transformers, the generic Segformer and UperNet-Swin models could map the date palm trees with an mFscore of 92% and an mIoU of 85.7%. The Segmenter model was the least accurate and achieved an mFscore ranging from 90.14% to 91.24% and an mIoU ranging from 82.6% to 84.4%. Figure 11 exemplifies the results of different semantic segmentation models for images that were selected from the multiscale testing

data. From left to right are the raw image, the ground-truth label, and the result of the Segmenter-ViT-small, the DPT-base, the UperNet-Swin-tiny, and the Segformer-MiT-b2.

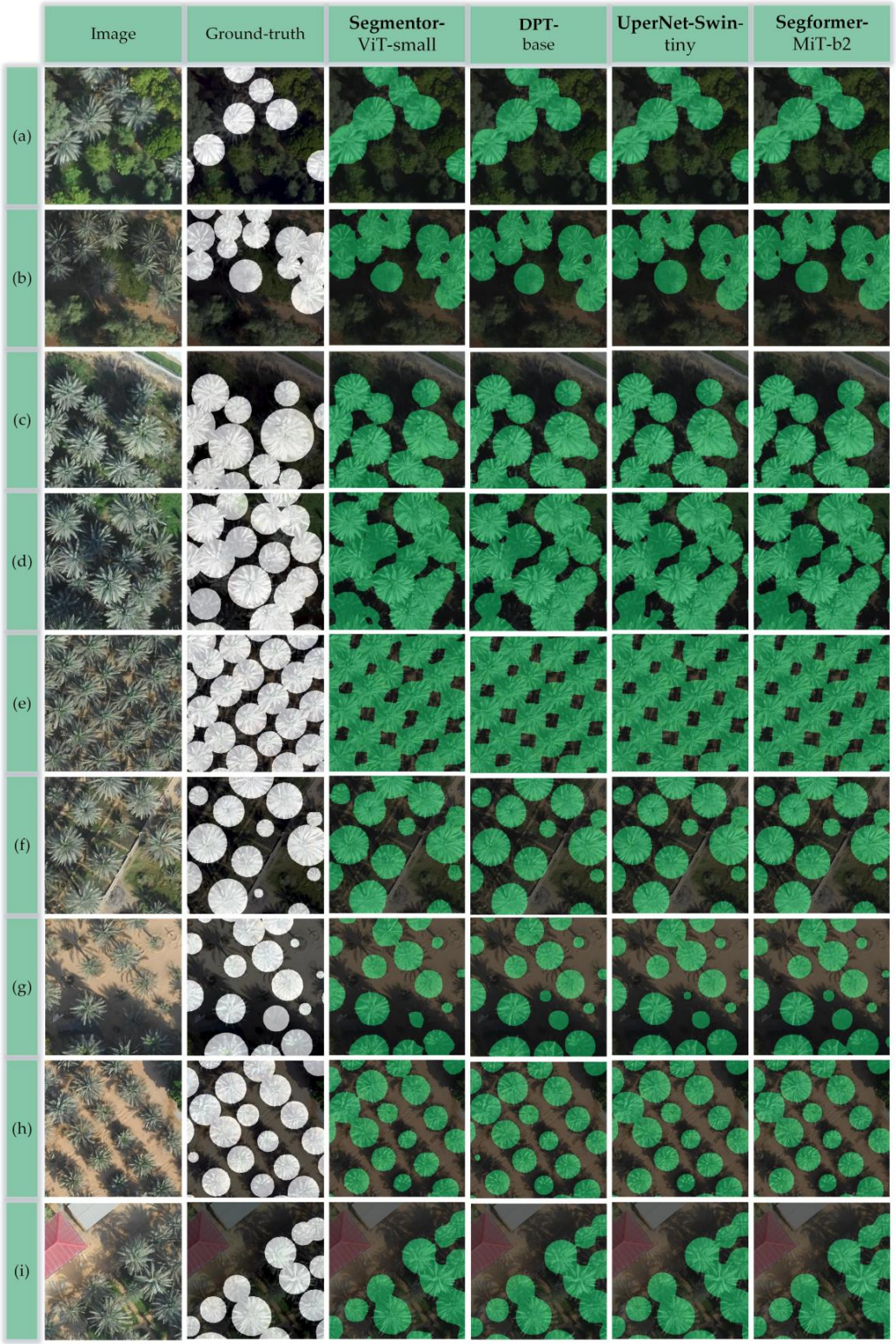

Figure 8. Segmentation results of date palm trees for a set of images (**a**–**i**) obtained from the UAV testing dataset. The left column lists the randomly selected images followed by the ground truth data and the results of Segmeter with ViT-small, DPT with ViT-base, UperNet with Swin-tiny, and Segformer with MiT-b2.

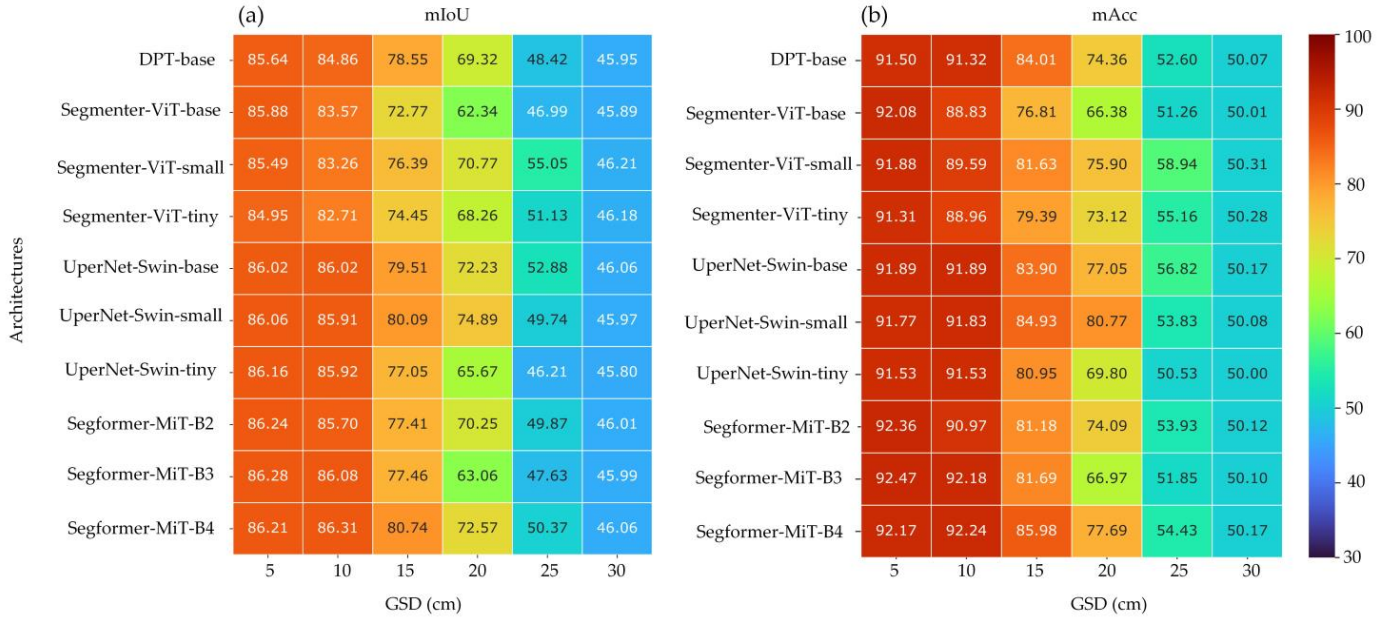

Figure 9. Results of generalizability assessment: (**a**) mIoU and (**b**) mAcc.

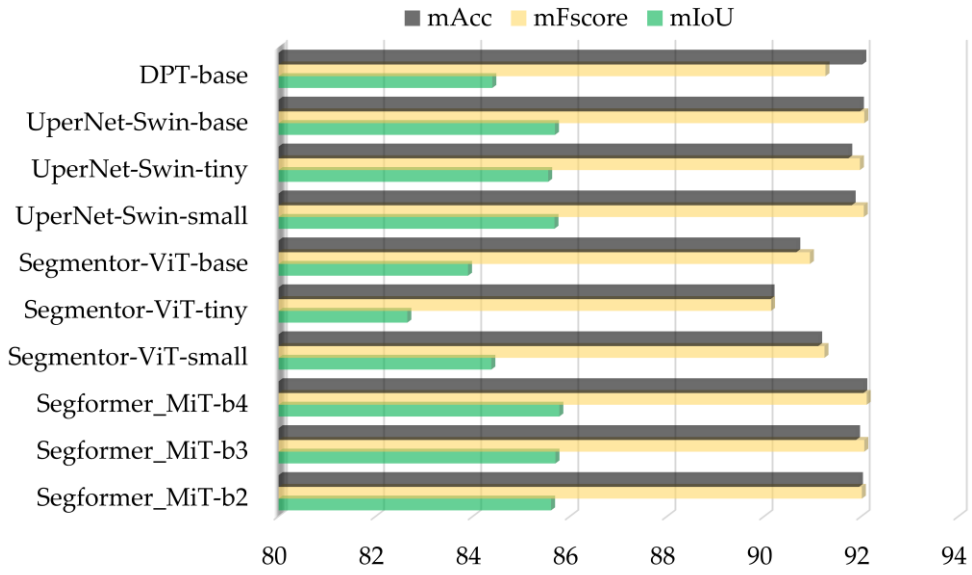

Figure 10. Segmentation accuracy metrics obtained from multiscale testing data.

4.3. Results of the Transferability Evaluation

The evaluated semantic segmentation models were first trained and evaluated on UAV data, then developed and assessed on multisource/multidate VHSR data. The UAV and multisource data were comprehensive and covered date palm trees with diverse characteristics (e.g., density, crown size, age, and height) and the surrounding environments (e.g., different vegetation, soil, and urban backgrounds). In order to ensure the transferability of the deep vision transformer to other VHSR images that were obtained by different sensors at other geographical locations, the semantic segmentation models were assessed on a public UAV dataset that was acquired in the Al Ain emirate [81]. Table 2 lists the transferability analysis of the models that were trained on the UAV-based data and the generic models that were trained on the combined multisource data. The results of the transferability analysis revealed that the evaluated models effectively delineated the date palm trees and achieved remarkable segmentation results from the unseen dataset. Both of

the models that were trained on the UAV and multiscale data almost maintained a similar range of accuracies. An example of the high degree of transferability to the unseen VHSR data is shown in Figure 12.

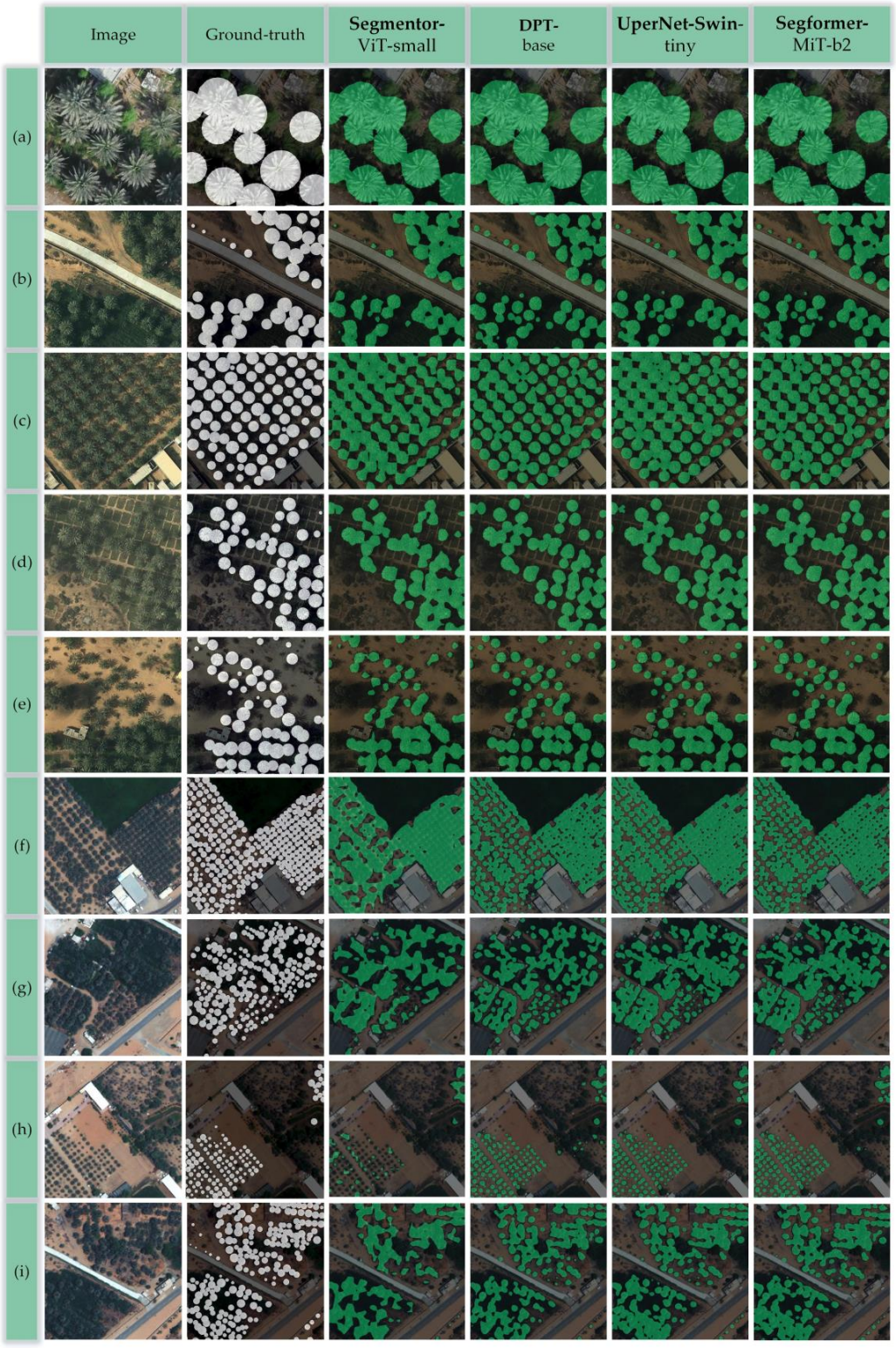

Figure 11. Segmentation results of date palm trees for a set of randomly selected multiscale images (**a**–**i**) chosen from the multiscale testing dataset. The left column lists the randomly selected images followed by the ground truth data and the results of the Segmenter-ViT-small, DPT-base, UperNet-Swin-tiny, and Segformer-MiT-b2.

Table 2. Quantitative transferability results of the segmentation models trained on UAV and multi-source data.

	UAV-Based Models			Generic Multisource Models		
	mIoU	**mF-Score**	**mAcc**	**mIoU**	**mF-Score**	**mAcc**
Segformer_MiT-b2	86	92.39	92.53	85.84	92.31	92.96
Segformer_MiT-b3	86.2	92.52	93.15	85.84	92.31	92.96
Segformer_MiT-b4	86.36	92.61	92.69	85.83	92.3	92.91
Segmenter-ViT-tiny	84.7	91.61	91.34	82.72	90.41	90.01
Segmenter-ViT-small	85.5	92.09	92.06	85.13	91.88	91.92
Segmenter-ViT-base	85.64	92.18	92.7	85.22	91.93	92.5
UperNet-Swin-tiny	85.57	92.14	92.51	85.14	91.88	91.62
UperNet-Swin-small	85.87	92.31	91.87	85.6	92.15	92.29
UperNet-Swin-base	85.9	92.34	92.82	85.73	92.24	92.63
DPT-base	85.45	92.06	91.69	84.92	91.76	92.37

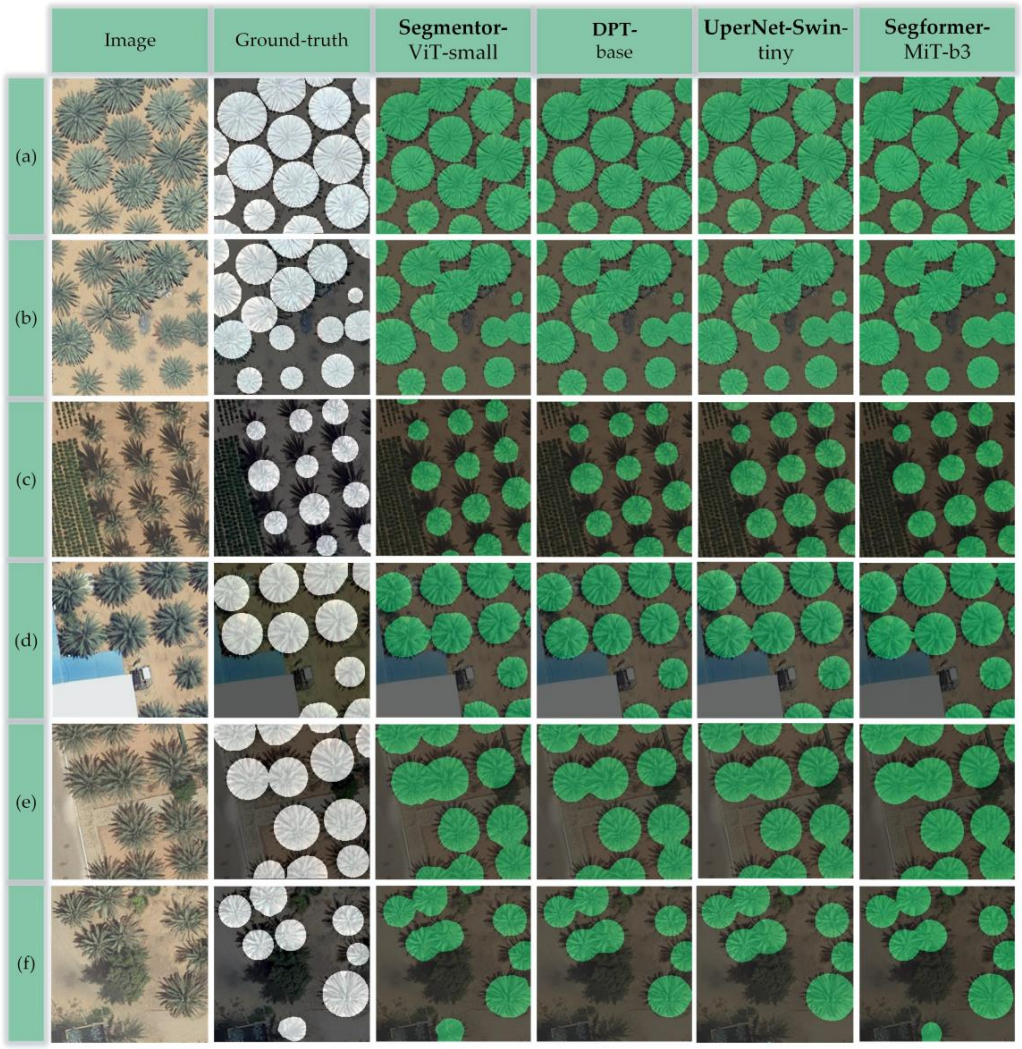

Figure 12. Segmentation results of date palm trees for a set of randomly selected images (**a**–**f**) chosen from the unseen dataset. The left column lists the randomly selected images, followed by the ground truth data and the results of the Segmenter-ViT-small, DPT-base, UperNet-Swin-tiny, and Segformer-MiT-b.

5. Discussion

The automatic development and update of date palm tree inventories are essential for their consistent monitoring, health and risk assessment, and sustainable management. In contrast to conventional field surveys, remote sensing provides an efficient means for observing and monitoring date palm trees at wide scales with very high resolutions. Deep learning techniques, especially CNNs, have shown remarkable performance in the extraction of date palm trees from UAV images through various object detection [83–85], semantic segmentation [36], and instance segmentation [68] models. The CNNs capture semantically rich information in the images through learnable layered convolutions. As the convolutional filters operate at the local level of the image, they hinder the capturing of global information and long-range semantic information interaction. Moreover, the details and the accuracy of CNN-based models are often reduced due to the loss of spatial information as a result of pooling/downsampling operations. However, some CNN-based techniques, such as DeepLab [34], can capture multiscale features by enlarging receptive fields through the utilization of dilated convolutions and the spatial pyramid pooling technique.

Given the limited spectral information of RGB images, the high intraclass variance of palm trees, the variations in context and background, and the differences in the spatial resolutions of the data, the accurate mapping of date palm trees from multisource and multidate data remains a challenge. In the current study, the evaluated deep vision transformers demonstrated remarkable performance in delineating date palm trees from UAV and multiscale VHSR images. The Segformer model, followed by the UperNet-Swin transformer model, achieved the highest segmentation results in both the UAV-based testing dataset (mF-score of 92.42% and mIoU of 86.3%) and the multiscale testing dataset (mF-score of 92.11% and mIoU of 85.8%). Figure 13 shows the results of the Segformer-MiT-b2 applied to a large subset that was extracted from the different multiscale images.

Prior to comparing the performance of the evaluated vision transformers with different CNN-based models, the FCN [86], PSPNet [35], DeepLab V3+ [34], and dual attention network for scene segmentation (DANet) [87] were trained and evaluated using multisource data. Moreover, the transferability of the CNN-based models to the unseen independent dataset was assessed. The residual learning network (ResNet-50) [88] was regarded as the backbone of the different CNN-based models. Figure 14a and b present the comparative performance of the transformer and the CNN-based models on the multiscale testing dataset and an unseen independent dataset, respectively. Figure 14c and d show a comparison of the number of parameters and the computational costs (in floating-point operations per second (FLOPs)) of the different models, respectively. Figure 14e and f display the training and the testing time of the transformer and the CNN-based models, respectively. The Segformer and UperNet-Swin models outperformed all of the evaluated CNN-based models on the multiscale and independent testing datasets. The Segmenter model, followed by the Segformer, entailed the lowest number of parameters, the lowest computational cost, and the least amount of training and inference time. Although the UperNet-Swin model showed excellent performance in mapping the date palm trees from different VHSR images, it was among the most computationally expensive models.

In general, the findings of this research are in line with those of several recent studies on remote sensing [89–92]. For instance, Tang et al. [89] employed Segformer (based on MiT-b4) to identify coseismic landslides from UAV images. The Segformer was compared with popular CNN-based semantic segmentation models, including Attention-Unet, U2Net, DeepLabV3, Fast-SCNN, and HRNet. The findings of their study demonstrated that the Segformer model was superior to the CNNs in landslide detection. Gonçalves et al. [91] compared the performances of the Segformer and the DPT with those of various CNN-based models (i.e., PSPNet, FCN, DeepLabV3+, OCRNet, and ISANet) in mapping the burned areas in the Brazilian Pantanal wetland from PlanetScope imagery. The Segformer model, followed by the DPT, outperforms the CNN-based models, and are the most accurate models in distinguishing the burned areas from the not-burned areas. In the present study,

the performance of the Segformer and the UperNet-Swin surpassed that of previous date palm tree mapping from UAV-based images, which was based on U-Net with a backbone of ResNet-50 [36]. The current study demonstrated an improvement in F-measure and mIoU of 1.4% and 1.3%, respectively.

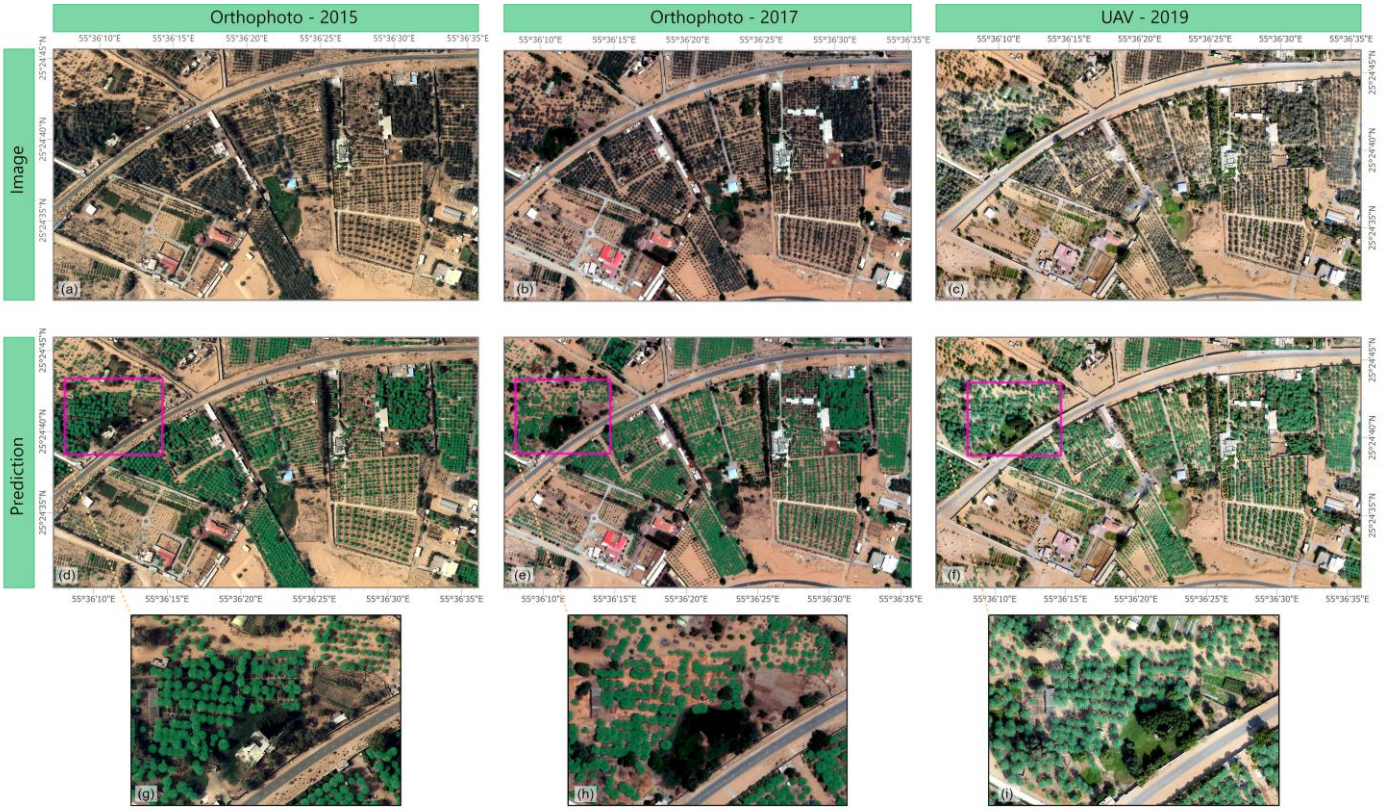

Figure 13. Segmentation results of Segformer-MiT-b2 applied to a large subset extracted from (**a**,**d**,**g**) an orthophoto with a GSD of 20 cm; (**b**,**e**,**h**) an orthophoto with a GSD of 15 cm, and (**c**,**f**,**i**) UAV data with a GSD of 5 cm.

Although the manual preparation of ground-truth (GT) data via visual assessment and radius delimitation procedures may sometimes result in creating boundaries that are slightly larger or smaller than the boundaries of the date palm trees, the actual boundaries of the date palm trees can be automatically learned by deep learning models regardless of the minor errors that may exist in the data. For instance, Figure 15a and b show the GT and the results that were obtained by the Segformer for the UAV image, and Figure 15c and d present the differences between the GT and the Segformer results for the aerial images. However, the discrepancy between the GT samples and the result of the segmentation models may consequently decrease the accuracy metrics. The Segformer model can effectively delineate the date palm trees that vary in size, shape, and density. Notably, the model is powerful and was able to delineate small date palm trees that had not been recognized by the analyst.

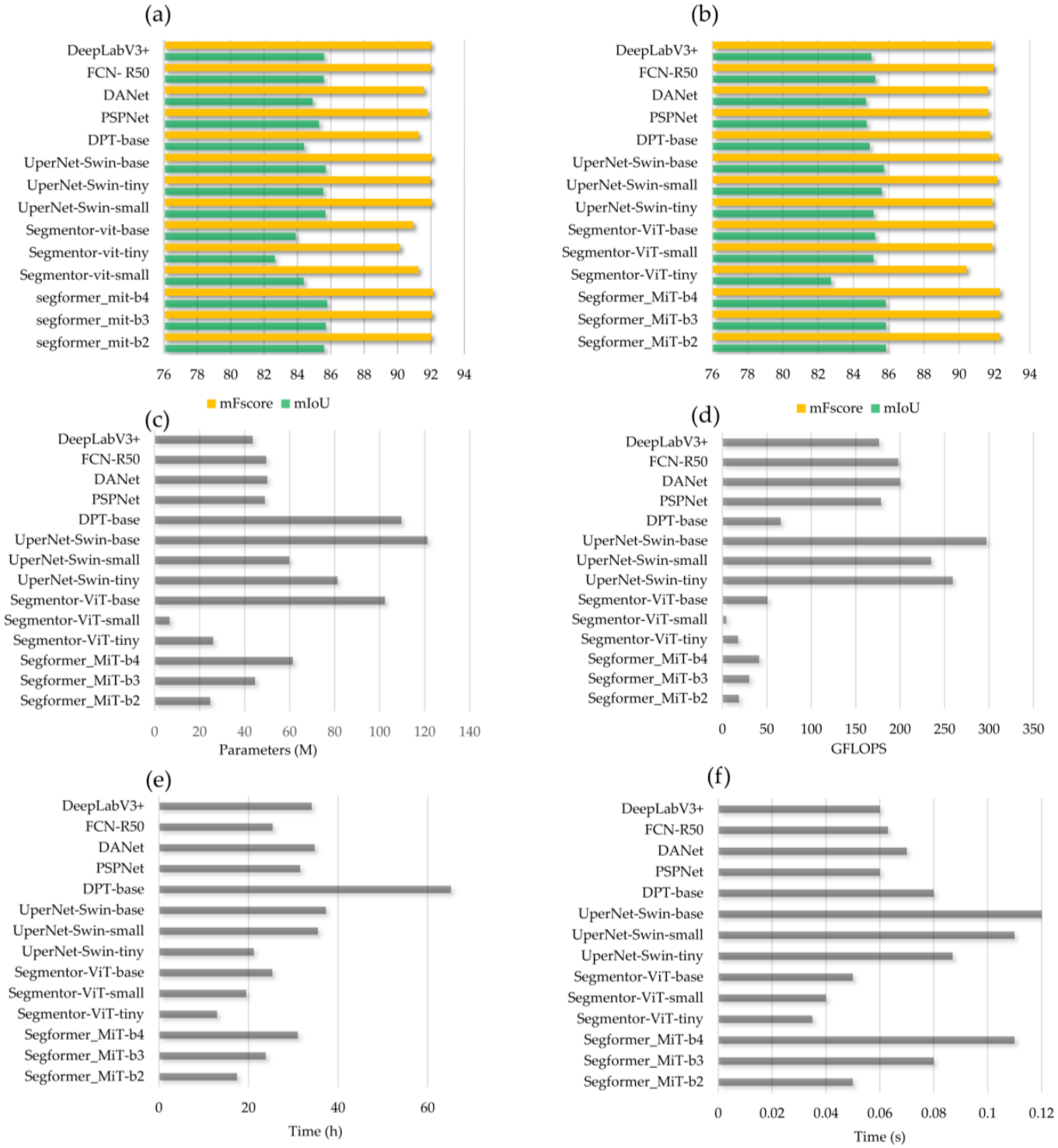

Figure 14. Comparisons of transformer-based and CNN-based semantic segmentation models in terms of (**a**) performance on multiscale testing data, (**b**) transferability to unseen data, (**c**) model parameters, (**d**) computational cost (in FLOPs), (**e**) training time, and (**f**) average inference time per image.

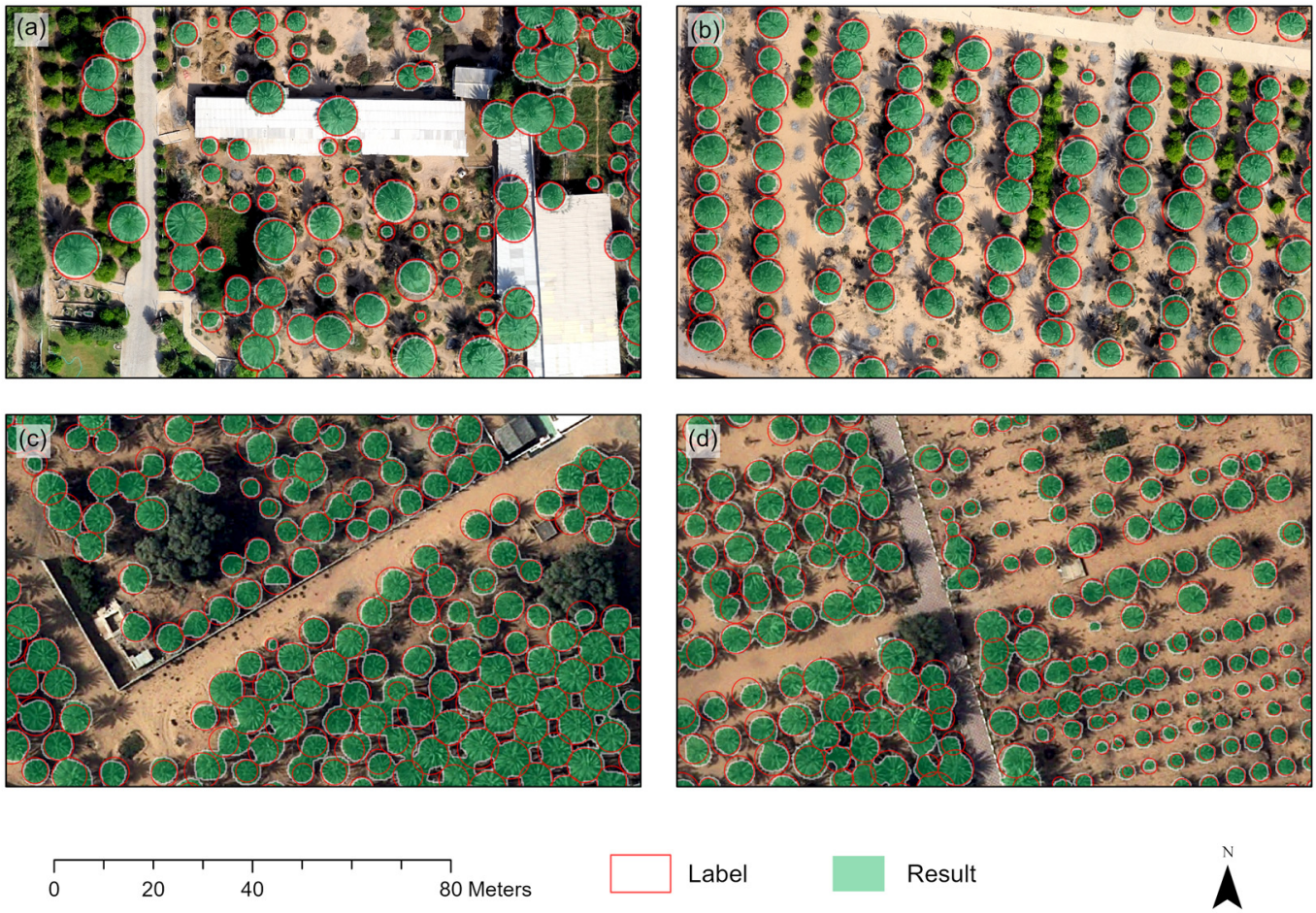

Figure 15. Potential discrepancies between the ground-truth labels and segmentation results in: (**a**,**b**) UAV, and (**c**,**d**) aerial images.

6. Conclusions

This study aimed to develop an end-to-end, reliable, and transferable deep learning approach for the large-scale mapping of date palm trees from multiscale UAV-based and multisource aerial images. The effectiveness of different deep vision transformers with various model complexities was explored and examined for the delineation of date palm trees, including the Segformer (MiT B2, B3, and B4), the Segmenter (ViT-tiny, ViT-small, and ViT-base), the UperNet-Swin transformer (tiny, small, and base), and the DPT (ViT-B). The semantic segmentation models were trained and evaluated on comprehensive UAV-based and multidate aerial images. The generalization capability of the evaluated vision transformers, which were developed on the UAV dataset, was examined on a multiscale UAV dataset (with spatial resolutions ranging from 5 cm to 30 cm). Moreover, the transferability of the vision transformer model to unseen UAV data that were acquired at different geographical regions was assessed and compared with that of several state-of-the-art CNN-based models.

The evaluated deep vision transformers showed promising results in mapping date palm trees from UAV images, with an mIoU ranging from 85% to 86.3% and an mF-score ranging from 91.62% to 92.44%. The Segformer model achieved the highest segmentation results in the UAV testing dataset (mF-score of 92.44% and mIoU of 86.3%) and the multi-scale testing dataset (mF-score of 92.11% and mIoU of 85.8%). Moreover, the Segformer model, followed by the UperNet-Swin transformer, surpassed all of the evaluated CNN-based models (including DeepLabV+, PSPNet, FCN-R50, and DANet) in the multiscale and the independent testing datasets. The Segformer model not only produced remarkable results in date palm tree mapping but was also among the models with a small number of

parameters and low computational cost. Overall, the performance of the evaluated deep vision transformers was comparable to the CNN-based models and demonstrated considerable generalization capabilities and transferability to different VHSR images. Deep vision transformers can efficiently be used for the accurate mapping and delineation of date palm trees from multidate and multiscale VHSR remotely sensed data, enabling the development and update of databases for different studies, as well as the consistent monitoring of date palm trees. Moreover, deep vision transformers can be efficiently utilized in segmentation tasks in versatile earth-related applications.

Author Contributions: Conceptualization, M.B.A.G., H.Z.M.S., R.A.-R. and A.S.; methodology, M.B.A.G., H.Z.M.S., R.A.-R., A.S. and F.N.; formal analysis, M.B.A.G.; data curation, M.B.A.G. and R.A.-R.; writing—original draft preparation, M.B.A.G.; writing—review and editing, M.B.A.G., H.Z.M.S., A.S., R.A.-R., F.N. and S.A.M.; visualization, M.B.A.G. All authors have read and agreed to the published version of the manuscript.

Funding: The project is funded by the University of Sharjah (UoS) under grant research project ID: 20020401163.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to acknowledge the University of Sharjah for the financial support and for providing the high-performance computing cluster that was used in this research, as well as the municipality of Ajman for providing remotely sensed data of the study area.

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

References

1. Zaid, A.; Wet, P.F. *Chapter I: Botanical and Systematic Description of the Date Palm*; FAO: Rome, Italy, 2002; Available online: http://www.fao.org/docrep/006.Y4360E/y4360e05.htm (accessed on 31 March 2018).
2. Spennemann, D.H.R. Review of the vertebrate-mediated dispersal of the Date Palm, *Phoenix dactylifera*. *Zool. Middle East* **2018**, *64*, 283–296. [CrossRef]
3. Krueger, R.R. Date palm (*Phoenix dactylifera* L.) biology and utilization. In *The Date Palm Genome*; Springer: Cham, Switzerland, 2021; Volume 1, pp. 3–28.
4. Food and Agriculture Organization. FAOSTAT. Available online: http://www.fao.org/faostat/en/#data/QC (accessed on 9 March 2021).
5. Mohan, M.; Silva, C.A.; Klauberg, C.; Jat, P.; Catts, G.; Cardil, A.; Hudak, A.T.; Dia, M. Individual tree detection from unmanned aerial vehicle (UAV) derived canopy height model in an open canopy mixed conifer forest. *Forests* **2017**, *8*, 340. [CrossRef]
6. Xi, X.; Xia, K.; Yang, Y.; Du, X.; Feng, H. Evaluation of dimensionality reduction methods for individual tree crown delineation using instance segmentation network and UAV multispectral imagery in urban forest. *Comput. Electron. Agric.* **2021**, *191*, 106506. [CrossRef]
7. Safonova, A.; Hamad, Y.; Dmitriev, E.; Georgiev, G.; Trenkin, V.; Georgieva, M.; Dimitrov, S.; Iliev, M. Individual tree crown delineation for the species classification and assessment of vital status of forest stands from UAV images. *Drones* **2021**, *5*, 77. [CrossRef]
8. Miraki, M.; Sohrabi, H.; Fatehi, P.; Kneubuehler, M. Individual tree crown delineation from high-resolution UAV images in broadleaf forest. *Ecol. Inform.* **2021**, *61*, 101207. [CrossRef]
9. Komárek, J.; Klápště, P.; Hrach, K.; Klouček, T. The Potential of Widespread UAV Cameras in the Identification of Conifers and the Delineation of Their Crowns. *Forests* **2022**, *13*, 710. [CrossRef]
10. Malek, S.; Bazi, Y.; Alajlan, N.; AlHichri, H.; Melgani, F. Efficient framework for palm tree detection in UAV images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 4692–4703. [CrossRef]
11. Chowdhury, P.N.; Shivakumara, P.; Nandanwar, L.; Samiron, F.; Pal, U.; Lu, T. Oil palm tree counting in drone images. *Pattern Recognit. Lett.* **2022**, *153*, 1–9. [CrossRef]
12. Han, P.; Ma, C.; Chen, J.; Chen, L.; Bu, S.; Xu, S.; Zhao, Y.; Zhang, C.; Hagino, T. Fast Tree Detection and Counting on UAVs for Sequential Aerial Images with Generating Orthophoto Mosaicing. *Remote Sens.* **2022**, *14*, 4113. [CrossRef]
13. Zhu, Y.; Zhou, J.; Yang, Y.; Liu, L.; Liu, F.; Kong, W. Rapid Target Detection of Fruit Trees Using UAV Imaging and Improved Light YOLOv4 Algorithm. *Remote Sens.* **2022**, *14*, 4324. [CrossRef]
14. Bazi, Y.; Malek, S.; Alajlan, N.; Alhichri, H. An automatic approach for palm tree counting in UAV images. In Proceedings of the 2014 IEEE Geoscience and Remote Sensing Symposium, Quebec City, QC, Canada, 13–18 July 2014; pp. 537–540.

15. Ecke, S.; Dempewolf, J.; Frey, J.; Schwaller, A.; Endres, E.; Klemmt, H.J.; Tiede, D.; Seifert, T. UAV-Based Forest Health Monitoring: A Systematic Review. *Remote Sens.* **2022**, *14*, 3205. [CrossRef]
16. Viera-Torres, M.; Sinde-González, I.; Gil-Docampo, M.; Bravo-Yandún, V.; Toulkeridis, T. Generating the baseline in the early detection of bud rot and red ring disease in oil palms by geospatial technologies. *Remote Sens.* **2020**, *12*, 3229. [CrossRef]
17. Li, W.; Dong, R.; Fu, H.; Yu, L. Large-scale oil palm tree detection from high-resolution satellite images using two-stage convolutional neural networks. *Remote Sens.* **2019**, *11*, 11. [CrossRef]
18. Hartling, S.; Sagan, V.; Sidike, P.; Maimaitijiang, M.; Carron, J. Urban tree species classification using a worldview-2/3 and liDAR data fusion approach and deep learning. *Sensors* **2019**, *19*, 1284. [CrossRef] [PubMed]
19. Pearse, G.D.; Watt, M.S.; Soewarto, J.; Tan, A.Y.S. Deep learning and phenology enhance large-scale tree species classification in aerial imagery during a biosecurity response. *Remote Sens.* **2021**, *13*, 1789. [CrossRef]
20. Kolanuvada, S.R.; Ilango, K.K. Automatic Extraction of Tree Crown for the Estimation of Biomass from UAV Imagery Using Neural Networks. *J. Indian Soc. Remote Sens.* **2021**, *49*, 651–658. [CrossRef]
21. Liu, X.; Ghazali, K.H.; Han, F.; Mohamed, I.I. Automatic Detection of Oil Palm Tree from UAV Images Based on the Deep Learning Method. *Appl. Artif. Intell.* **2021**, *35*, 13–24. [CrossRef]
22. Zamboni, P.; Junior, J.M.; Silva, J.d.A.; Miyoshi, G.T.; Matsubara, E.T.; Nogueira, K.; Gonçalves, W.N. Benchmarking anchor-based and anchor-free state-of-the-art deep learning methods for individual tree detection in rgb high-resolution images. *Remote Sens.* **2021**, *13*, 2482. [CrossRef]
23. Moura, M.M.; de Oliveira, L.E.S.; Sanquetta, C.R.; Bastos, A.; Mohan, M.; Corte, A.P.D. Towards Amazon Forest Restoration: Automatic Detection of Species from UAV Imagery. *Remote Sens.* **2021**, *13*, 2627. [CrossRef]
24. Xia, K.; Wang, H.; Yang, Y.; Du, X.; Feng, H. Automatic Detection and Parameter Estimation of Ginkgo biloba in Urban Environment Based on RGB Images. *J. Sens.* **2021**, *2021*, 668934. [CrossRef]
25. Veras, H.F.P.; Ferreira, M.P.; da Cunha Neto, E.M.; Figueiredo, E.O.; Corte, A.P.D.; Sanquetta, C.R. Fusing multi-season UAS images with convolutional neural networks to map tree species in Amazonian forests. *Ecol. Inform.* **2022**, *71*, 101815. [CrossRef]
26. Sun, Q.; Zhang, R.; Chen, L.; Zhang, L.; Zhang, H.; Zhao, C. Semantic segmentation and path planning for orchards based on UAV images. *Comput. Electron. Agric.* **2022**, *200*, 107222. [CrossRef]
27. Ji, Y.; Yan, E.; Yin, X.; Song, Y.; Wei, W.; Mo, D. Automated extraction of Camellia oleifera crown using unmanned aerial vehicle visible images and the ResU-Net deep learning model. *Front. Plant Sci.* **2022**, *13*, 958940. [CrossRef]
28. Lassalle, G.; Ferreira, M.P.; La Rosa, L.E.C.; de Souza Filho, C.R. Deep learning-based individual tree crown delineation in mangrove forests using very-high-resolution satellite imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *189*, 220–235. [CrossRef]
29. Zhang, C.; Zhou, J.; Wang, H.; Tan, T.; Cui, M.; Huang, Z.; Wang, P.; Zhang, L. Multi-Species Individual Tree Segmentation and Identification Based on Improved Mask R-CNN and UAV Imagery in Mixed Forests. *Remote Sens.* **2022**, *14*, 874. [CrossRef]
30. Yang, M.; Mou, Y.; Liu, S.; Meng, Y.; Liu, Z.; Li, P.; Xiang, W.; Zhou, X.; Peng, C. Detecting and mapping tree crowns based on convolutional neural network and Google Earth images. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *108*, 102764. [CrossRef]
31. Li, Y.; Chai, G.; Wang, Y.; Lei, L.; Zhang, X. ACE R-CNN: An Attention Complementary and Edge Detection-Based Instance Segmentation Algorithm for Individual Tree Species Identification Using UAV RGB Images and LiDAR Data. *Remote Sens.* **2022**, *14*, 3035. [CrossRef]
32. Badrinarayanan, V.; Handa, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. *arXiv* **2015**, arXiv:1505.07293.
33. Ronneberger, O.; Fischer, P.; Brox, T. 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)*; Springer: Berlin/Heidelberg, Germany, 2015; Volume 9351, pp. 234–241.
34. Chen, L.-C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking Atrous Convolution for Semantic Image Segmentation. *arXiv* **2017**, arXiv:1706.05587.
35. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, HI, USA, 21–26 July 2017; pp. 6230–6239.
36. Gibril, M.B.A.; Shafri, H.Z.M.; Shanableh, A.; Al-Ruzouq, R.; Wayayok, A.; Hashim, S.J. Deep convolutional neural network for large-scale date palm tree mapping from uav-based images. *Remote Sens.* **2021**, *13*, 2787. [CrossRef]
37. Anagnostis, A.; Tagarakis, A.C.; Kateris, D.; Moysiadis, V.; Sørensen, C.G.; Pearson, S.; Bochtis, D. Orchard Mapping with Deep Learning Semantic Segmentation. *Sensors* **2021**, *21*, 3813. [CrossRef] [PubMed]
38. Ferreira, M.P.; Lotte, R.G.; D'Elia, F.V.; Stamatopoulos, C.; Kim, D.H.; Benjamin, A.R. Accurate mapping of Brazil nut trees (Bertholletia excelsa) in Amazonian forests using WorldView-3 satellite images and convolutional neural networks. *Ecol. Inform.* **2021**, *63*, 101302. [CrossRef]
39. Freudenberg, M.; Nölke, N.; Agostini, A.; Urban, K.; Wörgötter, F.; Kleinn, C. Large scale palm tree detection in high resolution satellite images using U-Net. *Remote Sens.* **2019**, *11*, 312. [CrossRef]
40. Kentsch, S.; Caceres, M.L.L.; Serrano, D.; Roure, F.; Diez, Y. Computer vision and deep learning techniques for the analysis of drone-acquired forest images, a transfer learning study. *Remote Sens.* **2020**, *12*, 1287. [CrossRef]

41. Wagner, F.H.; Sanchez, A.; Tarabalka, Y.; Lotte, R.G.; Ferreira, M.P.; Aidar, M.P.M.; Gloor, E.; Phillips, O.L.; Aragão, L.E.O.C. Using the U-net convolutional network to map forest types and disturbance in the Atlantic rainforest with very high resolution images. *Remote Sens. Ecol. Conserv.* **2019**, *5*, 360–375. [CrossRef]
42. Wagner, F.H.; Sanchez, A.; Aidar, M.P.M.; Rochelle, A.L.C.; Tarabalka, Y.; Fonseca, M.G.; Phillips, O.L.; Gloor, E.; Aragão, L.E.O.C. Mapping Atlantic rainforest degradation and regeneration history with indicator species using convolutional network. *PLoS ONE* **2020**, *15*, e0229448. [CrossRef]
43. Liu, J.; Wang, X.; Wang, T. Classification of tree species and stock volume estimation in ground forest images using Deep Learning. *Comput. Electron. Agric.* **2019**, *166*, 105012. [CrossRef]
44. Kentsch, S.; Karatsiolis, S.; Kamilaris, A.; Tomhave, L.; Lopez Caceres, M.L. Identification of Tree Species in Japanese Forests based on Aerial Photography and Deep Learning. In *Advances and New Trends in Environmental Informatics*; Springer: Cham, Switzerland, 2020.
45. Schiefer, F.; Kattenborn, T.; Frick, A.; Frey, J.; Schall, P.; Koch, B.; Schmidtlein, S. Mapping forest tree species in high resolution UAV-based RGB-imagery by means of convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2020**, *170*, 205–215. [CrossRef]
46. Shang, G.; Liu, G.; Zhu, P.; Han, J.; Xia, C.; Jiang, K. A deep residual U-type network for semantic segmentation of orchard environments. *Appl. Sci.* **2021**, *11*, 322. [CrossRef]
47. Ayhan, B.; Kwan, C. Tree, shrub, and grass classification using only RGB images. *Remote Sens.* **2020**, *12*, 1333. [CrossRef]
48. Ferreira, M.P.; de Almeida, D.R.A.; Papa, D.d.A.; Minervino, J.B.S.; Veras, H.F.P.; Formighieri, A.; Santos, C.A.N.; Ferreira, M.A.D.; Figueiredo, E.O.; Ferreira, E.J.L. Individual tree detection and species classification of Amazonian palms using UAV images and deep learning. *For. Ecol. Manag.* **2020**, *475*, 118397. [CrossRef]
49. Cheng, Z.; Qi, L.; Cheng, Y. Cherry Tree Crown Extraction from Natural Orchard Images with Complex Backgrounds. *Agriculture* **2021**, *11*, 431. [CrossRef]
50. Morales, G.; Kemper, G.; Sevillano, G.; Arteaga, D.; Ortega, I.; Telles, J. Automatic segmentation of Mauritia flexuosa in unmanned aerial vehicle (UAV) imagery using deep learning. *Forests* **2018**, *9*, 736. [CrossRef]
51. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
52. Kolesnikov, A.; Dosovitskiy, A.; Weissenborn, D.; Heigold, G.; Uszkoreit, J.; Beyer, L.; Minderer, M.; Dehghani, M.; Houlsby, N.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2021**, arXiv:2010.11929.
53. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 213–229.
54. Dai, X.; Chen, Y.; Xiao, B.; Chen, D.; Liu, M.; Yuan, L.; Zhang, L. Dynamic head: Unifying object detection heads with attentions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021), Nashville, TN, USA, 20–25 June 2021; pp. 7373–7382.
55. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.S.; et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2021), Nashville, TN, USA, 20–25 June 2021; pp. 6881–6890.
56. Xia, Z.; Pan, X.; Song, S.; Li, L.E.; Huang, G. Vision Transformer with Deformable Attention. *arXiv* **2022**, arXiv:2201.00520.
57. Jamali, A.; Mahdianpari, M. Swin Transformer and Deep Convolutional Neural Networks for Coastal Wetland Classification Using Sentinel-1, Sentinel-2, and LiDAR Data. *Remote Sens.* **2022**, *14*, 359. [CrossRef]
58. Jamali, A.; Mahdianpari, M.; Brisco, B.; Mao, D.; Salehi, B.; Mohammadimanesh, F. 3DUNetGSFormer: A deep learning pipeline for complex wetland mapping using generative adversarial networks and Swin transformer. *Ecol. Inform.* **2022**, *72*, 101904. [CrossRef]
59. Mekhalfi, M.L.; Nicolo, C.; Bazi, Y.; Al Rahhal, M.M.; Alsharif, N.A.; Maghayreh, E. Al Contrasting YOLOv5, Transformer, and EfficientDet Detectors for Crop Circle Detection in Desert. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 19–23. [CrossRef]
60. Chen, G.; Shang, Y. Transformer for Tree Counting in Aerial Images. *Remote Sens.* **2022**, *14*, 476. [CrossRef]
61. Gao, L.; Liu, H.; Yang, M.; Chen, L.; Wan, Y.; Xiao, Z.; Qian, Y. STransFuse: Fusing Swin Transformer and Convolutional Neural Network for Remote Sensing Image Semantic Segmentation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 10990–11003. [CrossRef]
62. Zhang, C.; Wang, L.; Cheng, S.; Li, Y. SwinSUNet: Pure Transformer Network for Remote Sensing Image Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5224713. [CrossRef]
63. Chen, X.; Qiu, C.; Guo, W.; Yu, A.; Tong, X.; Schmitt, M. Multiscale Feature Learning by Transformer for Building Extraction from Satellite Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 2503605. [CrossRef]
64. Abozeid, A.; Alanazi, R.; Elhadad, A.; Taloba, A.I.; Abd El-Aziz, R.M. A Large-Scale Dataset and Deep Learning Model for Detecting and Counting Olive Trees in Satellite Imagery. *Comput. Intell. Neurosci.* **2022**, *2022*, 1549842. [CrossRef] [PubMed]
65. Yang, L.; Wang, X.; Zhai, J. Waterline Extraction for Artificial Coast with Vision Transformers. *Front. Environ. Sci.* **2022**, *10*, 799250. [CrossRef]

66. Panboonyuen, T.; Jitkajornwanich, K.; Lawawirojwong, S.; Srestasathiern, P.; Vateekul, P. Transformer-based decoder designs for semantic segmentation on remotely sensed images. *Remote Sens.* **2021**, *13*, 5100. [CrossRef]
67. Fan, F.; Zeng, X.; Wei, S.; Zhang, H.; Tang, D.; Shi, J.; Zhang, X. Efficient Instance Segmentation Paradigm for Interpreting SAR and Optical Images. *Remote Sens.* **2022**, *14*, 531. [CrossRef]
68. Gibril, M.B.A.; Shafri, H.Z.M.; Shanableh, A.; Al-Ruzouq, R.; Wayayok, A.; bin Hashim, S.J.; Sachit, M.S. Deep convolutional neural networks and Swin transformer-based frameworks for individual date palm tree detection and mapping from large-scale UAV images. *Geocarto Int.* **2021**, 1–31. [CrossRef]
69. Lan, Y.; Lin, S.; Du, H.; Guo, Y.; Deng, X. Real-Time UAV Patrol Technology in Orchard Based on the Swin-T YOLOX Lightweight Model. *Remote Sens.* **2022**, *14*, 5806. [CrossRef]
70. Alshammari, H.H.; Shahin, O.R. An Efficient Deep Learning Mechanism for the Recognition of Olive Trees in Jouf Region. *Comput. Intell. Neurosci.* **2022**, *2022*, 9249530. [CrossRef]
71. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *15*, 12077–12090.
72. Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. *arXiv* **2021**, arXiv:2102.12122.
73. Islam, M.A.; Jia, S.; Bruce, N.D.B. How Much Position Information Do Convolutional Neural Networks Encode? *arXiv* **2020**, arXiv:2001.08248.
74. Xiao, T.; Liu, Y.; Zhou, B.; Jiang, Y.; Sun, J. Unified Perceptual Parsing for Scene Understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Cham, Switzerland, 2018; pp. 432–448.
75. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
76. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 9992–10002.
77. Strudel, R.; Garcia, R.; Laptev, I.; Schmid, C. Segmenter: Transformer for Semantic Segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision; pp. 7242–7252. Available online: https://openaccess.thecvf.com/content/ICCV2021/papers/Strudel_Segmenter_Transformer_for_Semantic_Segmentation_ICCV_2021_paper.pdf (accessed on 6 December 2022).
78. Ranftl, R.; Bochkovskiy, A.; Koltun, V. Vision Transformers for Dense Prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision; pp. 12159–12168. Available online: https://openaccess.thecvf.com/content/ICCV2021/papers/Ranftl_Vision_Transformers_for_Dense_Prediction_ICCV_2021_paper.pdf (accessed on 6 December 2022).
79. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 8024–8035.
80. MMS Contributors. {MMSegmentation}: OpenMMLab Semantic Segmentation Toolbox and Benchmark. 2020. Available online: https://github.com/open-mmlab/mmsegmentation (accessed on 6 December 2022).
81. Al-Saad, M.; Aburaed, N.; Al Mansoori, S.; Ahmad, H. Al Autonomous Palm Tree Detection from Remote Sensing Images—UAE Dataset. In Proceedings of the IGARSS 2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; pp. 2191–2194.
82. Labelme/Examples/Semantic_Segmentation at Main Wkentaro/Labelme. Available online: https://github.com/wkentaro/labelme/tree/main/examples/semantic_segmentation (accessed on 6 December 2022).
83. Ammar, A.; Koubaa, A.; Benjdira, B. Deep-learning-based automated palm tree counting and geolocation in large farms from aerial geotagged images. *Agronomy* **2021**, *11*, 1458. [CrossRef]
84. Jintasuttisak, T.; Edirisinghe, E.; Elbattay, A. Deep neural network based date palm tree detection in drone imagery. *Comput. Electron. Agric.* **2022**, *192*, 106560. [CrossRef]
85. Culman, M.; Delalieux, S.; Van Tricht, K. Palm Tree Inventory From Aerial Images Using Retinanet. In Proceedings of the 2020 Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS), Tunis, Tunisia, 9–11 March 2020; pp. 314–317.
86. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
87. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2019), Long Beach, CA, USA, 15–20 June 2019; pp. 3141–3149.
88. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
89. Tang, X.; Tu, Z.; Wang, Y.; Liu, M.; Li, D.; Fan, X. Automatic Detection of Coseismic Landslides Using a New Transformer Method. *Remote Sens.* **2022**, *14*, 2884. [CrossRef]
90. Guo, F.; Qian, Y.; Liu, J.; Yu, H. Pavement crack detection based on transformer network. *Autom. Constr.* **2023**, *145*, 104646. [CrossRef]

91. Gonçalves, D.N.; Marcato, J.; Carrilho, A.C.; Acosta, P.R.; Ramos, A.P.M.; Gomes, F.D.G.; Osco, L.P.; da Rosa Oliveira, M.; Martins, J.A.C.; Damasceno, G.A.; et al. Transformers for mapping burned areas in Brazilian Pantanal and Amazon with PlanetScope imagery. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *116*, 103151. [CrossRef]
92. Jiang, K.; Afzaal, U.; Lee, J. Transformer-Based Weed Segmentation for Grass Management. *Sensors* **2023**, *23*, 65. [CrossRef] [PubMed]