A Comparative Analysis of Deep Learning Methods for Ghaf Tree Detection and Segmentation from UAV-based Images

Hani Shanableh [1], Mohamed Barakat A.Gibril [1,*], Ahmed Mansour [1], Aditya Dixit [1], Rami Al-Ruzouq [1,2], Nezar Hammouri [1], Fouad Lamghari [3], Safa M. Ahmed [1], Ratiranjan Jena [1], Tilal Mohamed [1], Mohammed Abdulraheem Almarzouqi [4], Nedal Salem Alafayfeh[4], Simon Zerisenay Ghebremeskel [3]

[1] GIS and Remote Sensing Center, Research Institute of Sciences and Engineering, University of Sharjah, Sharjah 27272, United Arab Emirates
[2] Civil and Environmental Engineering Department, University of Sharjah, Sharjah, Sharjah 27272, United Arab Emirates
[3] Fujairah Research Centre, Al-Hilal Tower, 3003, P.O. Box 666 Fujairah, United Arab Emirates
[4] UAE Ministry of Climate Change and Environment, United Arab Emirates

Keywords: Ghaf trees, Prosopis cineraria, tree crown delineation, instance segmentation, vision transformers.

Abstract

The *Prosopis cineraria*, commonly known as the Ghaf tree, is an ecologically significant species that prevents desertification, enhances soil fertility, and supports biodiversity within arid ecosystems. Mapping and monitoring Ghaf trees using unmanned aerial systems (UAVs) and deep learning are essential for advancing conservation efforts through reliable, automated assessments. In this study, we performed a comparative analysis of several transformer-based deep learning models, including Mask DETR with Improved Denoising Anchor Boxes (Mask DINO) and Mask R-CNN models based on the Vision Transformer (ViT), Swin Transformer, and Enhanced Multiscale ViT (MViTv2), for mapping Ghaf trees from UAV images captured in diverse urban and agricultural environments. Results demonstrated strong potential for the assessed instance segmentation architectures in mapping Ghaf trees, achieving mean average precision values of 80% to 84.2% for detection and 82.2% to 85.1% for segmentation, with F1-scores ranging from 83.55% to 88.3% for detection and 85.5% to 88.6% for segmentation. This study underscores the effectiveness of transformer-based deep learning architectures for mapping Ghaf trees from UAV images, with findings and refinements that can be applied and extended to map other native tree species and support broader conservation initiatives.

1. Introduction

The Ghaf tree (*Prosopis cineraria*) is often celebrated as the "tree of life" in regions such as Bahrain and parts of Arabia due to its extraordinary resilience in arid desert environments (Kalarikkal, Kim, and Ksiksi 2022). Known for its ability to survive centuries in harsh, hot climates without artificial irrigation, the Ghaf tree was declared the National Tree of the United Arab Emirates (UAE) in 2008, symbolizing its profound cultural and historical significance (Bhardwaj 2021). Traditionally, Ghaf leaves were used as fodder for camels, while young leaves continue to have culinary and medicinal applications. However, rapid urban development in the UAE has increasingly placed this iconic tree under threat (Gallacher and Hill 2005).

The Ghaf tree is ecologically critical, playing a pivotal role in preventing desertification, enhancing soil fertility, and supporting biodiversity in the UAE’s fragile desert ecosystem. This tree’s significant carbon sequestration capacity also aids in mitigating climate change impacts. In recognition of its ecological value, the UAE government has initiated conservation efforts like the "Give a Ghaf" campaign to replenish the tree’s population, now estimated at over 100,000 across the country (“The Ghaf Tree - National Tree of the UAE - Importance, Uses, Facts,”, 2004,) These initiatives highlight the Ghaf tree’s status as a symbol of sustainability within the UAE's environmental agenda.

Monitoring Ghaf habitats in the challenging desert landscape increasingly relies on aerial surveillance technologies, particularly Unmanned Aerial Vehicles (UAVs) (Sankey et al. 2018). UAV-based imagery has gained traction in recent years for a variety of ecological applications, including mapping woody plant encroachment in grasslands (Oddi et al. 2021) and estimating carbon stocks in desert vegetation (Abdullah, Al-Ali, and Srinivasan 2021). Despite its utility in large-scale, high-resolution mapping, real-time UAV application remains limited due to operational constraints (Hill and Rowan 2022; Hashemi-Beni et al. 2018).

In recent years, deep learning (DL) has emerged as the preferred approach for tackling complex computer vision (CV) tasks, including denoising, segmentation, and object detection, offering significant advantages over traditional machine learning (ML) methods (Dixit et al. 2023; Tang et al. 2023). This shift is largely due to DL's capacity for automatic feature extraction, in contrast to the manual feature engineering typical of classical ML. Convolutional Neural Networks (CNNs) are prominent DL techniques, especially effective in extracting features from imagery, making them ideal for segmentation tasks. Various CNN-based models have proven successful in identifying and mapping tree crowns using UAV imagery (Zhang et al. 2022; Li et al. 2022; Moura et al. 2021; Mo et al. 2021; Erdem et al. 2023; Sun et al. 2022).

Traditional CNNs focus on localized features through convolutional operations, which can restrict their ability to capture long-range dependencies (Dixit et al. 2023). To address this limitation, recent approaches have integrated self-attention (SA) mechanisms, enhancing CNNs' ability to capture global information (Yuan, Chen, and Wang 2020). Yet, many of these models still aggregate global context by combining localized feature maps rather than encoding it directly (Mou, Hua, and Zhu 2020). Consequently, fully leveraging CNNs for comprehensive context extraction in complex remote sensing tasks remains challenging.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-G-2025-805-2025

The advent of transformers (Vaswani 2017) has revolutionized CV by enabling models to capture global contextual information through parallelizable, order-independent operations. Vision transformers (ViTs) have demonstrated their strength in capturing global context, enhancing representational capacity in CV tasks (Dixit et al., 2023). In remote sensing, ViTs are increasingly investigated for their potential across various applications (Aleissaee et al., 2023; Dixit et al. 2023). Swin Transformers (ST) (Liu et al. 2021) further build on this capability by introducing a hierarchical structure with local SA within shifted windows, which allows for efficient computation and enhanced spatial localization. ST have already been explored as backbones in object detection and instance segmentation, including their integration with Mask R-CNN (He et al. 2017), where they have shown to improve precision by effectively capturing both local and global features (Gibril et al. 2024). This adaptability and efficiency make ST valuable in complex tasks like remote sensing and UAV-based image analysis, where accurate object delineation is critical.

This study aims to perform a comparative analysis of four DL-based models for the instance segmentation and mapping of Ghaf trees using UAV imagery. The contribution of this research is the evaluation of Mask R-CNN based on Swin Transformer, Mask DINO, and Mask R-CNN based ViT an enhanced multiscale ViT, to provide insights into accurate segmentation and delineation of Ghaf trees in challenging environments.

The paper is organized as follows: Section 2 describes the study area and dataset. Section 3 details the methodology, experimental setup, and evaluation metrics. Section 4 presents the results, and Section 5 summarizes the study's key findings and contributions.

2. Study area and dataset

2.1 Study Area

The study area, Figure 1, spans diverse urban and agricultural regions within the Fujairah and Sharjah emirates, focusing on some in Kalba and Fujairah city. The region covers an area of approximately 25 km² and includes various tree species, notably Ghaf, Neem, and *Acacia tortilis*. The selection of this study region allows for a comprehensive examination of vegetation across urban and agricultural settings within the Emirates.

For this study, imagery was collected using a Sensefly eBee X fixed-wing drone, equipped with a professional-grade Sensefly S.O.D.A RGB camera. Drone flights were conducted over the study area at a consistent altitude of 122 meters. Each flight adhered to civil aviation regulations, ensuring safe and efficient data capture across selected agricultural areas.

Flight missions were planned using eMotion mission planner software, establishing 40% vertical and 70% horizontal overlap to optimize coverage. The eBee X drone operates at speeds between 40 and 110 km/h and can manage winds up to 46 km/h, allowing for both belly landings and manual hand launches. With a single battery, the drone is capable of covering areas up to 5 square km, with additional extension options as needed.

The S.O.D.A camera onboard has a variable focal length (2.8–11 mm) and a 5,472 × 3,648 pixels resolution, capturing detailed imagery with a 3:2 aspect ratio. It offers exposure compensation of ±2.0 EV (in 1/3 EV increments), a global shutter speed ranging from 1/30 to 1/2000 seconds, and an ISO sensitivity between 125 and 6400. At the operational altitude of 122 meters (400 feet), the ground sampling distance of the generated orthomosaic is 2.5 cm per pixel, ensuring high-resolution data capture. Imagery was collected on clear days from 9:00 a.m. to 1:00 p.m. (GMT +3) for optimal lighting conditions.

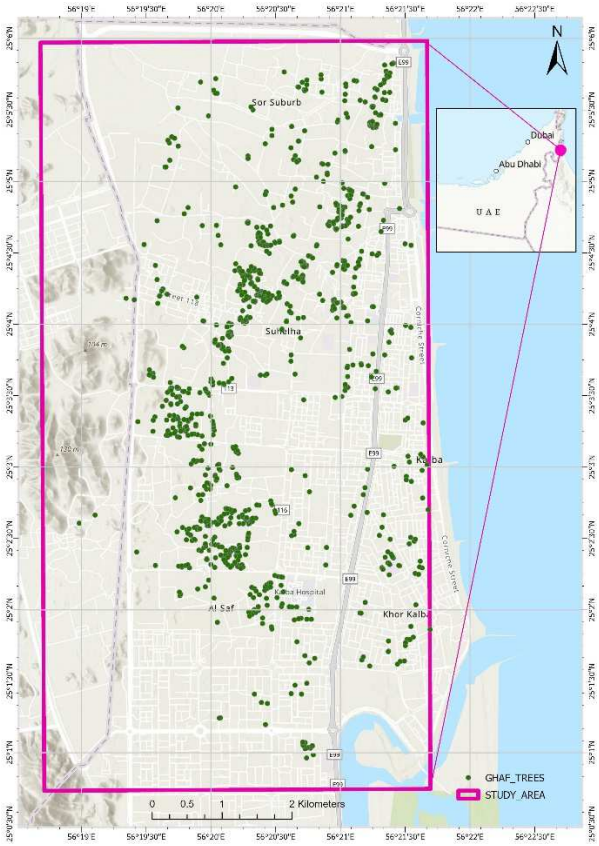

Figure 1. Geographic Map of Study Area.

2.2 Dataset

The preparation of input data for model training involved a detailed manual delineation of each Ghaf tree within the imagery. This process used field data alongside visual evaluations conducted through ArcGIS software to ensure accurate annotation. For the purpose of testing, validation, and training, the study area was segmented into three distinct regions, each providing UAV images and their corresponding labels. These images were then converted into annotation formats.

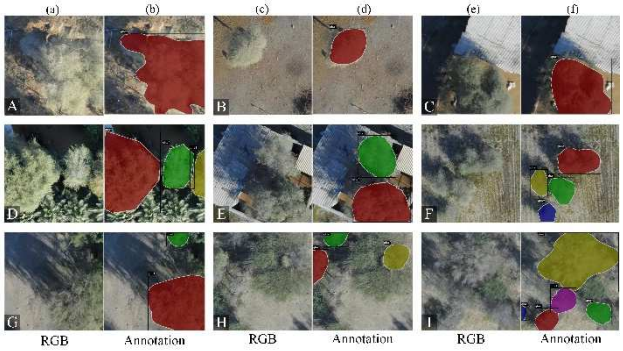

Figure 2. Representation of various image patches (A–I) with corresponding annotations.

In total, 2770 images in COCO format were used for training the evaluated DL models, encompassing 3890 instances of Ghaf trees. Additionally, 680 COCO-formatted images containing 1020 instances of Ghaf trees were reserved for validation. A distinct testing dataset was created to evaluate model performance, including 885 images with 1273 instances of Ghaf trees specifically set aside for this purpose. Annotated examples of Ghaf trees across image patches (A–I) are displayed in Figure 2.

3. Methodology

3.1 Instance Segment Architectures

In this work, we performed a comparative study on various transformer-based instance segmentation architectures to detect and map Ghaf trees from UAV data. The evaluated DL models include Mask R-CNN based on ViT, Swin transformer, and an enhanced multiscale ViT architecture. Moreover, the study included the analysis of Mask DINO (DETR with Improved Denoising Anchor Boxes).

Mask R-CNN is an extension of the Faster R-CNN object detection framework that adds a branch for predicting segmentation masks on a per-instance basis. It operates in two stages: the first stage generates region proposals, while the second stage classifies these proposals and refines their boundaries. In addition to bounding box regression and classification, Mask R-CNN introduces a pixel-level segmentation mask for each detected object, enabling precise object segmentation.

A Vision Transformer (ViT) (Dosovitskiy et al. 2021) is a DL model that uses a transformer-based architecture for image recognition. Unlike CNNs, which rely on convolutional layers, ViTs partition an image into fixed-size, non-overlapping patches and treat each patch as an individual "token." These patches are then flattened into vectors, with positional embeddings added to retain spatial context. The ViT architecture consists of multiple transformer layers, each containing self-attention (SA) mechanisms and feed-forward networks. The SA mechanism captures relationships between patches, enabling the model to learn long-range dependencies and better interpret complex visual patterns.

The multiscale ViT (Y. Li et al. 2022) uses a window attention mechanism to maintain tensor resolution while applying local SA, alongside pooled attention that aggregates features through downsampling and global SA. Traditional window attention only performs local SA within isolated windows, which limits connectivity across windows. To mitigate this, the multiscale ViT incorporates Hybrid Window Attention (Hwin), enabling cross-window connections by applying local attention within individual windows across most blocks, while the final three stages that connect to the Feature Pyramid Networks (FPN) (Lin et al. 2017) retain global context.

Swin Transformer (ST) (Liu et al. 2021) is a DL model that extends traditional transformers for image recognition by using a hierarchical structure and local SA within overlapping windows. Unlike ViTs, which use fixed-size patches, ST divides images into smaller windows, allowing efficient computation and better spatial localization. The architecture consists of stages that merge windows progressively, enabling the capture of both local and global features. The SA is computed within these windows, promoting computational efficiency, while shifted windows improve cross-window interactions for enhanced feature learning.

Mask DINO (F. Li et al. 2023) is a unified framework designed for object detection and segmentation tasks utilizing a transformer-based architecture. As an extension of DINO, Mask DINO incorporates a mask prediction branch capable of handling instance, panoptic, and semantic segmentation tasks, ensuring versatility across segmentation domains. It utilizes DINO's query embeddings to create high-resolution pixel embeddings, which enable the production of binary masks.

3.2 Experimental Settings

The experiments were conducted on a system equipped with 64 GB of RAM and an NVIDIA Titan RTX GPU. The models were initialized with pre-trained weights and trained over 100,000 iterations, with performance evaluations carried out every 2,000 iterations to identify the best-performing models for further analysis. The AdamW optimizer was utilized, set with an initial learning rate of 0.0001, and a batch size of 2 was employed throughout the training process.

3.3 Evaluation Metrics

In this study, we used the mean Average Precision (mAP) and F-score metrics to evaluate the performance. The Average Precision (AP) metric evaluates precision-recall (PR) performance for a specific class by calculating the area under the PR curve, which plots precision against recall across varying confidence thresholds. AP provides a score between 0 and 1, indicating the model's accuracy in correctly detecting instances of an object, with higher scores representing a better balance between precision and recall. The mAP and F-score metrics are calculated as follows:

$$P = \frac{TP}{TP+FP}, \quad (1)$$

$$R = \frac{TP}{TP+FN}, \quad (2)$$

$$IoU = \frac{\text{AoI}}{\text{AoU}}, \quad (3)$$

$$AP = \int P(r)dr, \quad (4)$$

$$F - score = \frac{2\times P\times R}{P+R}, \quad (5)$$

where P = Precision; R = Recall;
TP = true positive;
FP = false positive;
FN = false negative;
AoI = Area of intersection;
AoU = Area of union.

4. Results

The Mask R-CNN models showed notable differences in training times, with the MViTv2-tiny backbone completing the process in the shortest time at 6 hours, followed by the Swin transformer backbone at 10 hours, and the ViT-base backbone at 17.6 hours. Mask DINO with Swin transformer backbone required the longest time at 40 hours. Each model was evaluated on the validation dataset every 2000 iterations, and the best-performing weights were selected for further processing. The detection and segmentation results on the validation and testing datasets are presented in Table 1.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-G-2025-805-2025 | © Author(s) 2025. CC BY 4.0 License.

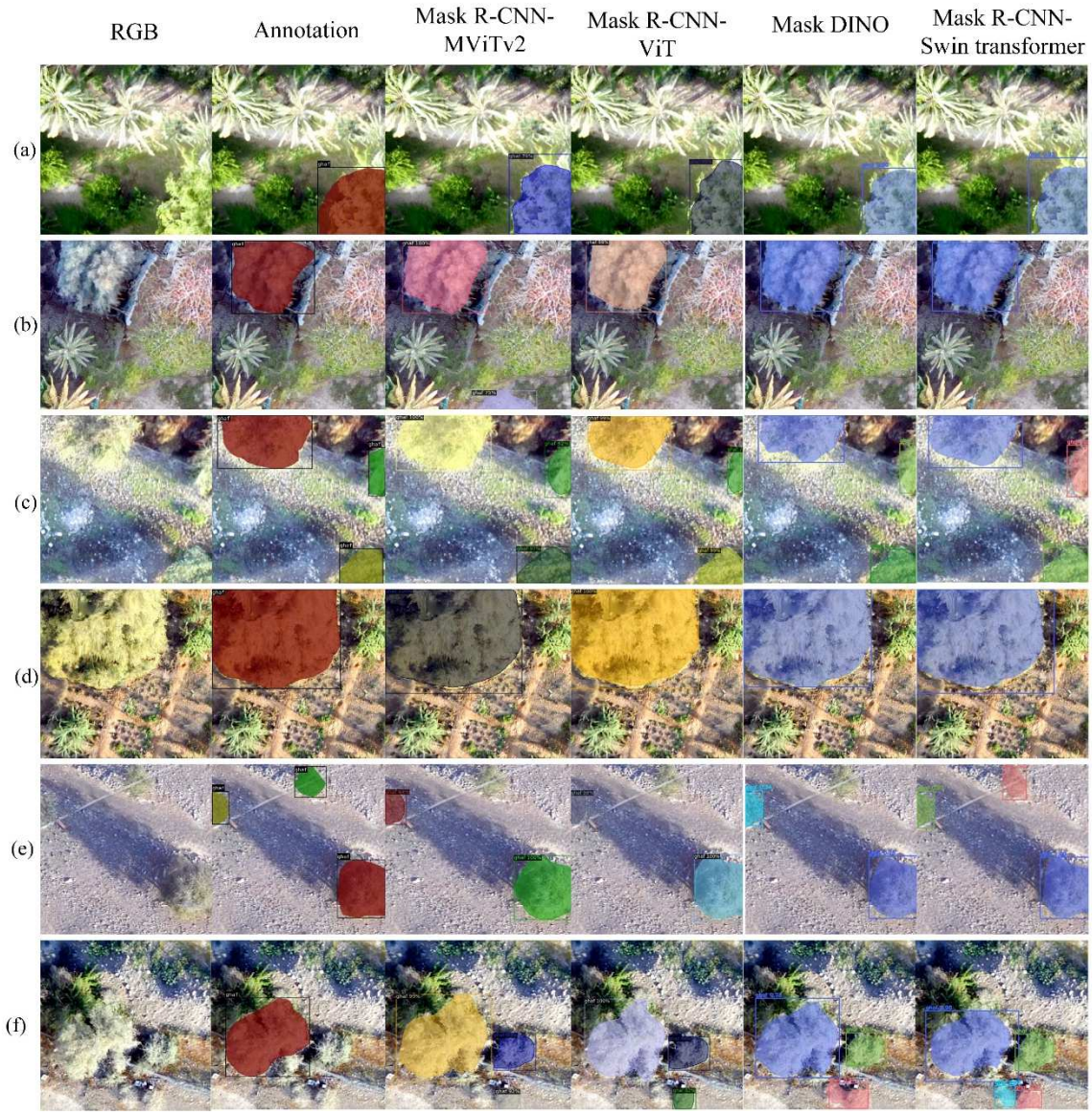

Figure 3. Selected images from the testing dataset (a–f) with corresponding annotations, presenting results from four models: Mask R-CNN with MViTv2-tiny Mask R-CNN with ViT-base, Mask DINO, and Mask R-CNN with Swin transformer.

		Detection		**Segmentation**	
		mAP_{50}	F1-score	mAP_{50}	F1-score
Validation	Mask R-CNN-Swin transformer	82.0	86.1	82.6	86.4
	Mask DINO-Swin-tiny	76.0	79.3	76.4	79.5
	Mask R-CNN-MViT-tiny	80.1	83.9	79.6	83.5
	Mask R-CNN-ViT-base	80.5	84.8	79.5	83.7
Testing	Mask R-CNN-Swin transformer	84.2	88.3	85.1	88.6
	Mask DINO-Swin-tiny	84.1	87	84.8	87.6
	Mask R-CNN-MViTv2-tiny	80.0	83.55	82.2	85.5
	Mask R-CNN-ViT-base	83.8	86.4	84.0	86.49

Table 1. Experimental results of the assessed DL network

The mAP_{50} values for Acacia tree detection on the validation set ranged from 76% to 82%, while segmentation mAP_{50} scores ranged from 76.4% to 82.6%. For the F1-score, detection values on the validation set varied between 79.3% and 86.1%, and segmentation scores ranged from 79.5% to 86.4%.

The evaluation on the testing dataset showed mAP_{50} values for Acacia tree detection between 80% and 84.2%, with segmentation mAP_{50} values ranging from 82.2% to 85.1%. Similarly, F1-scores for detection varied from 83.55% to 88.3%, while segmentation F1-scores ranged from 85.5% to 88.6%. The Mask R-CNN model with the Swin Transformer backbone surpassed the other evaluated models in detecting and segmenting Ghaf trees on the testing dataset, achieving an mAP of 84.2% and an F1-score of 88.3%. Although the Mask DINO model with the Swin Transformer backbone showed lower performance on the validation data, it reached a mAP_{50} of 84.1 and 84.8% and an F1-score of 87.0 and 87.6 on the testing dataset for detection and segmentation tasks, respectively. The Mask R-CNN models with ViT-base and MViTv2 backbones demonstrated competitive segmentation results, achieving F1 scores of 86.14% and 85.5%, respectively.

https://doi.org/10.5194/isprs-annals-X-G-2025-805-2025

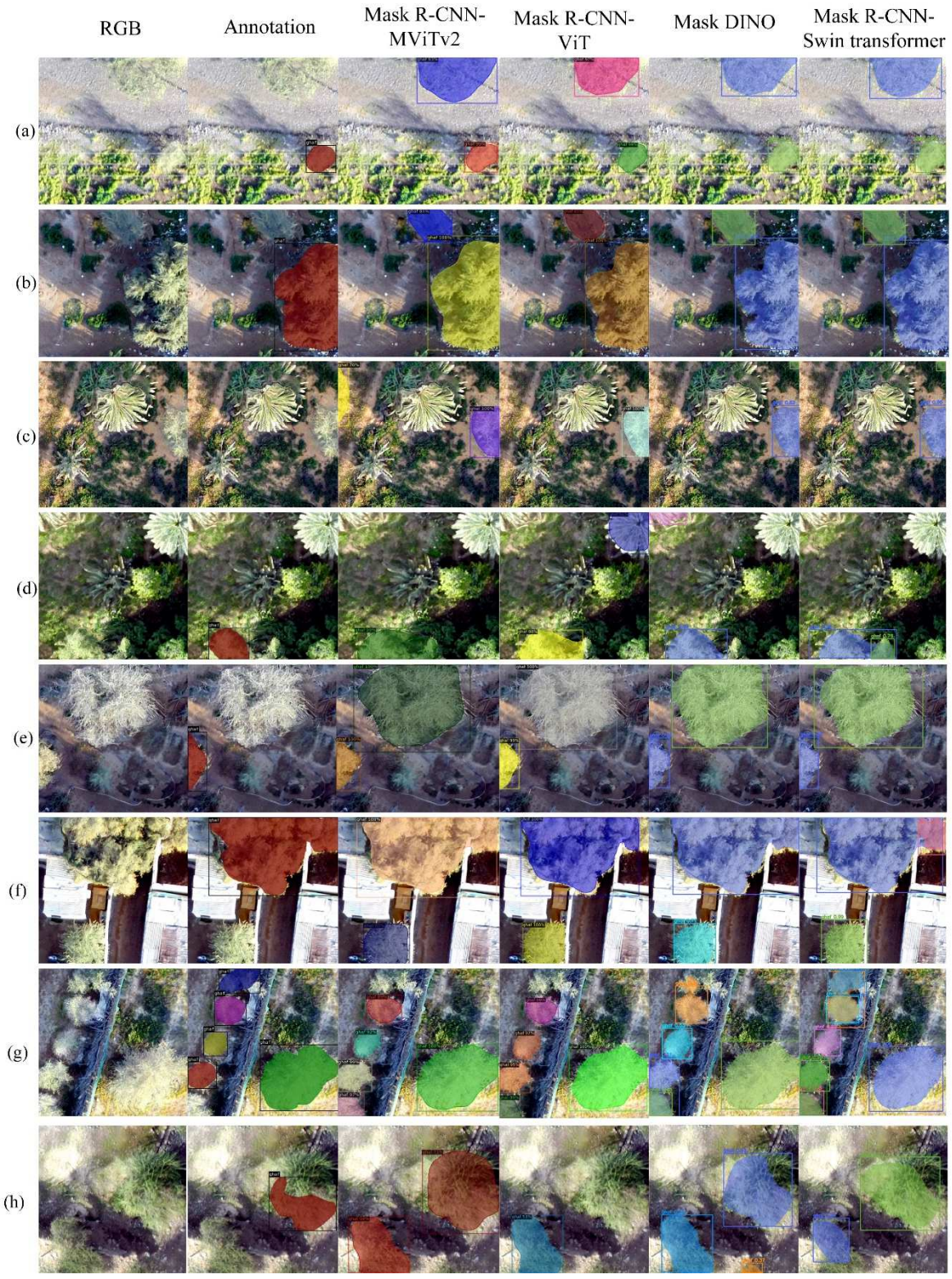

Figure 4. Examples illustrating the performance of the evaluated models in mapping Ghaf trees across diverse and challenging scenes.

The evaluated models demonstrated promising capabilities in mapping Ghaf trees across diverse urban and agricultural landscapes, yet several classification challenges persist. Figure 4 illustrates various cases where the evaluated models overlooked or misclassified Ghaf trees. The quality of acquiring and preprocessing UAV data can influence model accuracy in recognizing Ghaf trees. For instance, windy conditions during UAV data acquisition can lead to a loss of structural detail in Ghaf trees (Figures 4 a and b), which is essential for distinguishing tree species. This blurring or distortion often obscures fine leaf features, causing some trees, such as Acacia trees, to appear fluffy, clouded, and rounded, with branches

grouped, resembling Ghaf trees' structure. Transformers rely on extracted global and contextual information—such as tree shape, shadows, and surrounding vegetation—to accurately recognize Ghaf trees. However, splitting the UAV data into smaller image tiles (e.g., 512 × 512) can result in losing important context, especially for trees positioned at tile edges, which might lead to overlooking or misclassifying Ghaf trees, as shown in Figure 4c. Under certain lighting conditions, intense reflections can cause some trees to appear similarly bright to Ghaf trees, leading to potential misclassifications (Figure 4d). One of the primary challenges arises from the difficulty in acquiring a large, diverse ground-truth dataset across the vast and heterogeneous areas where these trees are found, along with potential errors introduced during tree count augmentation through image interpretation of RGB data. Including ambiguous cases in the training data can reinforce misclassifications, complicating accurate between similar tree species. Addressing these complexities is essential to enhance deep learning models for accurate, scalable mapping of this important tree species across varied environments.

5. Conclusion

Preserving Ghaf trees is essential for combating desertification and supporting biodiversity within desert ecosystems. This study was motivated by the need for precise mapping and monitoring of Ghaf trees using unmanned aerial systems (UAVs) and deep learning, aiming to enhance conservation efforts through reliable, automated assessments. A comparative analysis was conducted on several transformer-based deep learning models, including Mask DETR with Improved Denoising Anchor Boxes and Mask R-CNN models based on the Vision Transformer (ViT), Swin Transformer, and Enhanced Multiscale ViT. These models leverage the capability of vision transformers to capture global and contextual information, thereby enhancing mapping accuracy. The results showed that the evaluated instance segmentation architectures hold strong potential for mapping Ghaf trees, with mean average precision values between 80% and 84.2% for detection and 82.2% to 85.1% for segmentation, and F1-scores ranging from 83.55% to 88.3% for detection and 85.5% to 88.6% for segmentation.. The findings of this research revealed that variations in UAV image quality—caused by environmental factors such as wind-induced motion blur and shifting shadow positions—introduce challenges that affect the model's precision and recall. Specifically, these conditions can lead the model to misclassify non-Ghaf trees as Ghaf trees or Ghaf trees as non-Ghaf trees. However, despite these challenges, the evaluated transformer-based models demonstrated strong potential for real-world applications, highlighting its robustness in diverse environmental conditions. This study demonstrates the effectiveness of transformer-based deep learning architectures in mapping Ghaf trees from UAV images, offering insights and advancements that can be adapted to map other native tree species and support broader conservation efforts.

References

Abdullah, Meshal M., Zahraa M. Al-Ali, and Shruthi Srinivasan. 2021. "The Use of UAV-Based Remote Sensing to Estimate Biomass and Carbon Stock for Native Desert Shrubs." *MethodsX* 8 (January):101399.

Aleissaee, Abdulaziz Amer, Amandeep Kumar, Rao Muhammad Anwer, Salman Khan, Hisham Cholakkal, Gui-Song Xia, and Fahad Shahbaz Khan. 2023. "Transformers in Remote Sensing: A Survey." *Remote Sensing* 15 (7): 1860.

BHARDWAJ, VIBHA. 2021. "PODS OF Prosopis Cineraria (GHAF): A GIFT OF NATURE FOR NUTRACEUTICAL." *Journal of Global Ecology and Environment* 11 (1): 15–18.

Dixit, Aditya, Anup Kumar Gupta, Puneet Gupta, Saurabh Srivastava, and Ankur Garg. 2023. "UNFOLD: 3D U-Net, 3D CNN and 3D Transformer Based Hyperspectral Image Denoising." *IEEE Transactions on Geoscience and Remote Sensing*.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, et al. 2021. "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale." arXiv. https://doi.org/10.48550/arXiv.2010.11929.

Erdem, Firat, Nuri Erkin Ocer, Dilek Kucuk Matci, Gordana Kaplan, and Ugur Avdan. 2023. "Apricot Tree Detection from UAV-Images Using Mask R-CNN and U-Net." *Photogrammetric Engineering & Remote Sensing* 89 (2): 89–96.

Gallacher, David, and Jeffrey Hill. 2005. "Status of Prosopis Cineraria (Ghaf) Tree Clusters in the Dubai Desert Conservation Reserve." *Tribulus* 15 (2).

Gibril, Mohamed Barakat A., Helmi Zulhaidi Mohd Shafri, Abdallah Shanableh, Rami Al-Ruzouq, Aimrun Wayayok, Shaiful Jahari bin Hashim, and Mourtadha Sarhan Sachit. 2022. "Deep Convolutional Neural Networks and Swin Transformer-Based Frameworks for Individual Date Palm Tree Detection and Mapping from Large-Scale UAV Images." *Geocarto International* 37 (27): 18569–99.

Gibril, Mohamed Barakat A., Rami Al-Ruzouq, Jan Bolcek, Abdallah Shanableh, and Ratiranjan Jena. 2024. "Building Extraction from Satellite Images Using Mask R-CNN and Swin Transformer." In *2024 34th International Conference Radioelektronika (RADIOELEKTRONIKA)*, 1–5. IEEE.

Hashemi-Beni, Leila, Jeffery Jones, Gary Thompson, Curt Johnson, and Asmamaw Gebrehiwot. 2018. "Challenges and Opportunities for UAV-Based Digital Elevation Model Generation for Flood-Risk Management: A Case of Princeville, North Carolina." *Sensors* 18 (11): 3843.

He, Kaiming, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. "Mask R-Cnn." In *Proceedings of the IEEE International Conference on Computer Vision*, 2961–69.

Hill, Austin Chad, and Yorke M. Rowan. 2022. "The Black Desert Drone Survey: New Perspectives on an Ancient Landscape." *Remote Sensing* 14 (3): 702.

Kalarikkal, Remya Kottarathu, Youngwook Kim, and Taoufik Ksiksi. 2022. "Incorporating Satellite Remote Sensing for Improving Potential Habitat Simulation of *Prosopis Cineraria* (L.) Druce in United Arab Emirates." *Global Ecology and Conservation* 37 (September):e02167.

Li, Feng, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M. Ni, and Heung-Yeung Shum. 2023. "Mask Dino: Towards a Unified Transformer-Based Framework for Object Detection and Segmentation." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3041–50.

Li, Weijia, Haohuan Fu, Le Yu, and Arthur Cracknell. 2016. "Deep Learning Based Oil Palm Tree Detection and Counting for High-Resolution Remote Sensing Images." *Remote Sensing* 9 (1): 22.

Li, Yanghao, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. 2022. “MViTv2: Improved Multiscale Vision Transformers for Classification and Detection.” In Proceedings of the *IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 4804–14.

Li, Yingbo, Guoqi Chai, Yueting Wang, Lingting Lei, and Xiaoli Zhang. 2022. “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 Sensing* 14 (13): 3035.

Lin, Tsung-Yi, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. “Feature Pyramid Networks for Object Detection.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2117–25.

Liu, Ze, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. “Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows.” In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–22.

Mo, Jiawei, Yubin Lan, Dongzi Yang, Fei Wen, Hongbin Qiu, Xin Chen, and Xiaoling Deng. 2021. “Deep Learning-Based Instance Segmentation Method of Litchi Canopy from UAV-Acquired Images.” *Remote Sensing* 13 (19): 3919.

Mou, Lichao, Yuansheng Hua, and Xiao Xiang Zhu. 2020. “Relation Matters: Relational Context-Aware Fully Convolutional Network for Semantic Segmentation of High-Resolution Aerial Images.” *IEEE Transactions on Geoscience and Remote Sensing* 58 (11): 7557–69.

Moura, Marks Melo, Luiz Eduardo Soares de Oliveira, Carlos Roberto Sanquetta, Alexis Bastos, Midhun Mohan, and Ana Paula Dalla Corte. 2021. “Towards Amazon Forest Restoration: Automatic Detection of Species from UAV Imagery.” *Remote Sensing* 13 (13): 2627.

Oddi, Ludovica, Edoardo Cremonese, Lorenzo Ascari, Gianluca Filippa, Marta Galvagno, Davide Serafino, and Umberto Morra di Cella. 2021. “Using UAV Imagery to Detect and Map Woody Species Encroachment in a Subalpine Grassland: Advantages and Limits.” *Remote Sensing* 13 (7): 1239.

Sankey, Temuulen T., Jason McVay, Tyson L. Swetnam, Mitchel P. McClaran, Philip Heilman, and Mary Nichols. 2018. “UAV Hyperspectral and Lidar Data and Their Fusion for Arid and Semi-arid Land Vegetation Monitoring.” Edited by Nathalie Pettorelli and Ned Horning. *Remote Sensing in Ecology and Conservation* 4 (1): 20–33.

Sun, Ying, Ziming Li, Huagui He, Liang Guo, Xinchang Zhang, and Qinchuan Xin. 2022. “Counting Trees in a Subtropical Mega City Using the Instance Segmentation Method.” *International Journal of Applied Earth Observation and Geoinformation* 106 (February):102662.

Tang, Guangyi, Jianjun Ni, Yonghao Zhao, Yang Gu, and Weidong Cao. 2023. “A Survey of Object Detection for UAVs Based on Deep Learning.” *Remote Sensing* 16 (1): 149.

“The Ghaf Tree - National Tree Of The UAE - Importance, Uses, Facts.” n.d. Accessed October 10, 2024. https://www.dubai-online.com/essential/national-tree-uae/.

Vaswani, A. 2017. “Attention Is All You Need.” *Advances in Neural Information Processing Systems*. https://docalysis.com/files/hpgzy/download/attention%20all%20you%20need.pdf.

Yuan, Yuhui, Xilin Chen, and Jingdong Wang. 2020. “Object-Contextual Representations for Semantic Segmentation.” In *Computer Vision – ECCV 2020*, edited by Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, 12351:173–90. Lecture Notes in Computer Science. Cham: Springer International Publishing.

Zhang, Chong, Jiawei Zhou, Huiwen Wang, Tianyi Tan, Mengchen Cui, Zilu Huang, Pei Wang, and Li Zhang. 2022. “Multi-Species Individual Tree Segmentation and Identification Based on Improved Mask R-CNN and UAV Imagery in Mixed Forests.” *Remote Sensing* 14 (4): 874.