Comments Reviewer #1			
ID	**Line**	**Comment**	**Response**
1	184-188	Other than learning rate, batch size, and epoch, did you tune other parameters? Also, for learning rate, batch size, and epoch, it is better to test with a wider range of values to evaluate model performance before narrowing them down to a specific range. Also, for model training, did you use k-fold cross-validation for hyperparameter tuning? If so, what is the k-fold value did you use? This needs to be clarified.	We thank the reviewer for pointing this out. Yes, we tested different hyperparameter settings both in this study and in our earlier work (Soltani et al., 2024). The reported parameter settings yielded improved results and were therefore adopted. In the revised version, we will describe the hyperparameters and how we selected them in more depth. We did not use k-fold cross-validation. The models were evaluated on an entirely independent test dataset (see Section 2.1.1).
2	239-243	The prediction of acquisition distance seems skeptical. In citizen science data, people use various cameras and may set various zooming modes when capturing photos, it is hard to predict acquisition distance just from the photo itself; thus, distance thresholds of 0.2 m and 20 m seem skeptical. In the earlier paragraph, authors use an area threshold of 30% to filter out some photos. Should a similar method be used to filter out photos with large amounts of tree trunk/branch?	We understand the reviewer's skepticism about estimating camera-to-plant distance from a single photo. Inferring absolute distance is indeed challenging without known camera parameters. Our approach was intended to exclude extremely close-up photos showing individual leaves, or very distant photos showing broad landscapes. It does not aim to provide precise distance estimations but rather to filter out these two extreme cases. The applied threshold effectively removed such images, allowing us to include photos taken at distances commonly found in close-range UAV imagery. We will include visual examples in the supplementary information to transparently demonstrate the effectiveness of this method, which was already successfully applied in Soltani et al. (2019).
3	278-284	Did you use k-fold cross-validation to train the model? If so, the k-fold value you used should be reported.	We did not use k-fold cross-validation during model training. Final model evaluation was performed using manually delineated reference data from UAV images that were

			completely excluded from the training process (see Section 2.1.1). We will make this clearer in the revised manuscript.
4	286-301	The classification performance seems to be low for various species. Citizen science data helps reduce time and labor in reference data collection; however, we also need to make sure output data are accurate and usable. With this low accuracy, what do authors suggest for future works? Should we incorporate some UAV-based high accuracy labelled data in the model together with citizen science data to improve classification accuracy? Also, the hyperparameter tuning seems not to be well-performed in your deep learning model training, I recommend conducting a more exhaustive tuning and trying different deep learning architecture to see if the classification results are improved	We acknowledge the reviewer's concerns regarding the partially moderate segmentation accuracy and appreciate the forward-looking suggestions. First of all, we would like to highlight that using citizen science data for drone-based remote sensing is still in its infancy, and we are just pioneering the possibilities. This study is not about providing an operational technology, but rather about exploring methodological ways to harness citizen science data and its potential for drone-based mapping. Here, we demonstrate this potential in a very complex scenario, where several broadleaved tree species with very similar leaf forms are present. Given this pioneering character and the complexity of the case study, we are of the opinion that the results are groundbreaking and open up possibilities for a series of follow-up studies. Clearly, there are many aspects that can be improved and explored in greater depth (see also our Outlook section in the Discussion). In the revised manuscript, we will make it clearer that this study is of a pioneering nature and focuses on method development rather than providing a ready-to-use solution. We explored several strategies to improve segmentation accuracy across all tree species, including data augmentation, modifications of photograph backgrounds and scaling, hyperparameter tuning, and adjustments to model architectures. However, visual similarities among certain species led to trade-offs, improving accuracy for one species sometimes decreased it for others.

			Over several months, we conducted a thorough model ablation study, and the results presented here are the final outcome. In the revised manuscript, we will provide more information on these model ablations.
5		One of the main reasons that cause low segmentation accuracy in this study could be the difference in the spatial resolutions between citizen science photos and UAV images. One possible solution for this discrepancy could be that during your segmentation model training, authors may want to manipulate/resample citizen science photos to different resolutions, including the 0.22 cm resolution of the UAV image, and incorporate features extracted from these layers into the final segmentation prediction to help improve the final segmentation results (see below paper with similar idea, note: this is not a reviewer's paper). Martins et al., 2020. Exploring multiscale object-based convolutional neural network (multi-OCNN) for remote sensing image classification at high spatial resolution. https://doi.org/10.1016/j.isprsjprs.2020.08.004	We would first like to point out that for several species in the study, we have very high segmentation accuracy and overall model performance (e.g., F1 score above 0.5 for *Acer pseudoplatanus*, *Tilia platyphyllos, Quercus petraea,* and *Carpinus betulus*). This is particularly striking given that we did not acquire any specific training data and that the approach is entirely based on crowd-sourcing. It is important to note that this study is primarily about providing a methodological framework and showcasing the potential of such an approach. It is expected that this method will not (yet) work for all species, as this is truly pioneering work. In fact, we also aim to highlight the limitations of the approach, for instance, where it does not work well, such as with species that have very similar leaves (see Discussion, lines 377–404). Accordingly, in the revised manuscript, we will describe the overall objectives and limitations of this study more clearly. We agree that differences in spatial resolution and perspective present a challenge for our transfer learning approach. In our current implementation, we addressed this in part by downscaling, duplicating, and zooming out the citizen science photos before using them for training (see Section 2.3). This increased the likelihood that the appearance of plant

			features in citizen science imagery and drone-based imagery would align. Achieving a perfect resolution match is difficult due to variability in ground-level photo distances, image quality, and variation in the drone-based imagery (e.g., due to differences in canopy height). Instead, we applied a generic scaling strategy to reduce the level of detail across all ground photographs. This, combined with data augmentation, helped the model learn more scale-invariant features, which in turn improved generalization to UAV-scale imagery. Thank you for providing the reference on multiscale architectures. From our experience across a range of projects, we found that standard ("vanilla") architectures can learn multiscale phenomena on the fly when sufficient variability is present (see above), and when the model is deep enough. However, testing multiscale models in more depth is certainly promising, particularly if depth information is available, and we will include a discussion of this in the Outlook section of the revised version.