Comments Reviewer #2			
ID	**Line**	**Comment**	**Response**
1		I would recommend the authors to add a workflow chart to help readers understand the various types of methods and data used for the study. There are several AI/ML models employed for various different data processing, including both photographs and UAV imagery. I found it hard to connect the different processing steps, and how different data streams and AI/ML methods are used.	Thank you for this feedback. A flowchart will certainly help guide the reader through the methodological approach. We will therefore add a workflow chart in the revised version.
2		Second, not much information is presented in the Results, barely enough to understand the performance of the model. The authors did quite significant work on processing and segmenting the photographs from iNaturealist and Pl@ntNet. However, results about these processing and segmentation are completely missed in the Results. I am nervous the presentation of Results is disconnected with the Methods. Recommend the authors to carefully tie them together, especially, how F1 score, confusion matrix was calculated. The authors mentioned independent transect validation data were identified from UAV imagery, but did not mention where and how those were produced, distribution across species and space etc. I think it is also useful to present the species maps across the experiment plots.	Thank you for this very important feedback! We acknowledge that the original Results section was too brief and appreciate the reviewer's suggestions. We will revise the section to include the missing details and ensure stronger alignment with the Methods section. In particular, we will elaborate on the description of the independent test data creation (transects). We will include an explanation of how the F1 score and confusion matrix were calculated. Additionally, we will create maps displaying the predictions versus the transect reference data. Together with the flowchart (suggested by the reviewer; see comment no. 1), we assume that the results and workflow will become more comprehensive.
3		Lastly an overall thought, a core advance of using UAV imagery is to provide landscape-scale observations. The authors argued that ultra-high (finer than 0.22 cm)	We agree that canopy structure and form could carry interesting information for species recognition. However, its value is very limited. For instance, in Schiefer et al. (2020), we

		might be necessary to better segment species from UAV imagery. This statement appears to "false", and ignored that canopy structure and form are important information for species identification, which are not considered in this study. On the other hand, it is cool to generate the initial masks for UAV species identification using photographs, but it might be more useful to iterate over the species segmentation at UAV level, leveraging other information like canopy form and structure, to enlarge training samples at UAV level, instead of forcing UAV data to the same resolution as ground photographs?	found a clear trend that higher spatial resolution imagery (leaf features) clearly outperforms coarser imagery (only including canopy structure) (https://doi.org/10.1016/j.isprsjprs.2020.10.015). This was also confirmed by various studies (see our review in Kattenborn et al., 2021, https://doi.org/10.1016/j.isprsjprs.2020.12.010). Thus, it can be assumed that detailed features, such as leaf forms, are the cardinal features to differentiate most species. This aligns well with species-recognition protocols used by humans. However, note that we do not specifically ignore canopy structure. Many citizen science observations do show entire plants and their canopies. This information is, in fact, included in the model. It seems that we did not clearly visualize and communicate the variability in the underlying citizen science imagery. Therefore, in the revised manuscript, we will include a figure in the supplementary information with several examples per species, showing the large variability in the images, including both leaf-level and whole-plant-level photographs. From this, it should become evident that the model could integrate features of the canopy structure if important. Thank you for highlighting this shortcoming!
4		I wonder what features the authors used for segmentation? It is clear that the authors used only RGB imagery, but are other indices or transformations incorporated in the SAM segmentation?	The reviewer is correct that we used only RGB image data for both the SAM-based segmentation and the CNN model. We did not compute or incorporate any additional spectral or textural indices, as SAM is primarily trained on RGB imagery and hence, any sophisticated feature engineering does not appear to be necessary. This is confirmed by the overall high quality

			of the output (see Fig. 3). In the revised manuscript, we will make it clear that we used only RGB imagery as input.
5		The author mentioned that photos/masks from citizen science were ‘zoomed’ out when applied as training for UAV imagery. What’s the resolution after that? Is it comparable to UAV resolution?	Since the original citizen science images had varying resolutions, applying fixed factors for zooming-out did not result in a uniform output resolution. However, our aim was to approximate the UAV resolution as closely as possible. We tested several factors for the zoom-out, and the selected value provided the best model performance . We thank the reviewer for this observation and will include a description of the resolution of the zoomed-out photos and masks in the revised manuscript.