Enhancing aerial imagery analysis: leveraging explainability and segmentation

Anany Dwivedi
AI Institute, University of Waikato
Hamilton, New Zealand
anany.dwivedi@waikato.ac.nz,

Nick Lim
AI Institute, University of Waikato
Hamilton, New Zealand
nick.lim@waikato.ac.nz,

Albert Bifet
AI Institute, University of Waikato
Hamilton, New Zealand
albert.bifet@waikato.ac.nz,

Eibe Frank
AI Institute, University of Waikato
Hamilton, New Zealand
eibe.frank@waikato.ac.nz,

Bernhard Pfahringer
AI Institute, University of Waikato
Hamilton, New Zealand
bernhard@waikato.ac.nz,

***Abstract*—In the field of aerial and satellite remote sensing, the widespread adoption of deep learning brings new possibilities. Current approaches, however, often overlook the unique characteristics of aerial data. This study introduces a methodology that capitalizes on distinctive features, leveraging additional annotations for enhanced neural network training. Despite modest gains in classification accuracy, the synergy of enhanced explainability, automated segmentation, and targeted classification demonstrates nuanced improvements. Preliminary results showcase potential applications in land cover mapping. This work can be extented towards reducing dependency on labor-intensive human annotations through an iterative annotation and training loop.**

***Index Terms*—Deep Learning, Aerial Imagery, Remote Sensing, Classification**

I. Introduction

In the recent years, the proliferation of deep learning methods into the field of aerial imagery and satellite-based remote sensing has provided practitioners new tools in land cover and land use mapping, crop and hazards monitoring, land-cover change detection, as well as opened possibilities for novel applications like super-resolution [1]–[5]. While the application of deep learning in this domain is becoming more common [6], the current approaches to deep learning, particularly in classification and segmentation of aerial and satellite images still use conventional image classification approaches designed for natural images.

The conventional approaches do not take in account some of the distinctive characteristics inherent to aerial imagery data, namely the predominantly overhead and (almost) orthorectified views, and the availability of additional annotations such as the cover-type polygons provided by geo-spatial authorities. In this research, we propose methodology that harnesses the power of these additional annotations for improved neural network training, and demonstrate the potential of leveraging on the stronger supervision to improve explanation quality and classification accuracy.

This work was supported by the TAIAO project, funded by the Ministry of Business, Innovation, and Employment (MBIE).

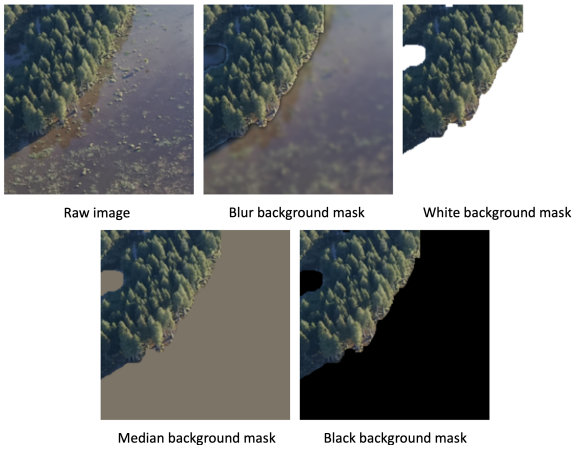

Fig. 1. Example of the types of image masking used.

Additionally, work by [7] shows that there is a symbiotic relationship between explanation quality and the overall classification accuracy. In this work, we extend this concept by providing preliminary results on how the improved explanations can be used as seed points for automated segmentation techniques such as Segment Anything Model (SAM) [8] and samgeo [9] to improve on manual segmentation.

II. Datasets

For developing and testing the proposed framework, we employ two different datasets. The first dataset contains images for Kahikatea (an indigenous tree in New Zealand) classification [10]. The task is constructed as a binary classification task, with positive class denoted as kahikatea trees present in the image, and negative as kahikatea trees not present in the image. The positive examples are annotated with the segmentation explanations provided by expert annotators.

The second dataset we use is the Land Cover Database (LCDB) and Waikato Regional Aerial Photography paired imagery for the Waikato region taken in 2019 [11], [12]. In this dataset, the polygons for a particular land cover type were

extracted from LCDB and the accompanying aerial images from the Waikato Regional Aerial Photography database to create the explanations. LCDB has 33 different land cover types, which were grouped into 7 super classes [13].

III. Methods

In this work, we use ResNet18 architecture which was pretrained on the imagenet dataset [14] to develop the classification models. The two datasets discussed in section II were used to train models and evaluate models separately. These classification models are trained and validated in two sets. In the first set, the models were trained on the raw images from the dataset. This was used as the baseline. In the second set, an ensemble of two models was created, where one model was the one obtained in the first set and the second model was obtained by fine-tuning the first model by masking the non-relevant land cover type from images in the dataset. To create the masked images, the image explanation was taken and the background was replaced by either a white mask, black mask, a median mask of the imagenet dataset (as the ResNet used was pretrained on imagenet), or with a gaussian blur mask added on the original background in that image. Examples of this type of image masking are shown in Fig. 1. The cross-entropy loss was used to train both the classification models. The model trained on raw images focuses on achieving high classification accuracy by minimizing the cross-entropy loss between the predicted and actual labels. In contrast, the second model is trained to teach the model focus on the area of interest through the use of masked images. To compute ensemble predictions, the logits outputted by the two models are combined using the formula $\mathcal{L} = \alpha\mathcal{L}_{\text{raw}} + (1-\alpha)\mathcal{L}_{\text{masked}}$, where $\alpha \in [0, 1]$, $\mathcal{L}_{\text{raw}}$ and $\mathcal{L}_{\text{masked}}$ are the vectors of logits from the models trained on raw images and masked images respectively. We evaluate the performance of the models based on the classification accuracy on the raw images in the test set.

IV. Results

Table I shows the classification accuracies obtained on both datasets tested in this work. For each dataset, the classification accuracy for all the image sets (raw images and four types of masked images) is presented. It can be noticed that the ensemble models perform better than the baseline (α = 1.0) in case of the Kahikatea dataset, showing a $\sim 5\%$ increase, but in case of the aerial photography dataset, the ensemble of the models performs similarly to the baseline (see Table I). Furthermore, for a mixing ratio of $\alpha = 0.9$, the ensemble models have similar or better classification accuracy performance than the baseline. Further research is needed to identify the main reason for this observation.

V. Discussion and Conclusion

We observed that using ensemble methods that utilize models trained using raw images and masked images, improves the accuracy of the classification. Moreover, we also observed improvements in the model explainability as measured by the Jaccard similarity between the Grad-CAM explainations [15] to the ground truth segmentation masks (not included here due to space constraints).

TABLE I
Classification accuracy for the two datasets and five different types of image inputs used for training the models. All the accuracies are in %s.

Dataset	α	Blur	White	Median	Black
Kahikatea	0.8	76.53	73.34	73.43	74.23
	0.9	78.31	78.69	**82.14**	78.95
	Baseline	76.91			
Aerial	0.8	72.63	72.42	72.62	72.83
	0.9	73.07	73.12	73.19	**73.22**
	Baseline	73.12			

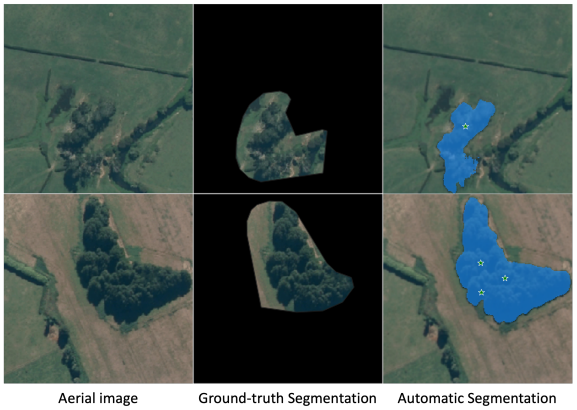

Fig. 2. Examples of improved segmentation masks through automatic segmentations using Grad-CAM explanations as seed points for SAM.

It is instructive to consider whether we can potentially leverage segmentations provided by SAM by using the coordinates of the top Grad-CAM explanations from the explanation-focused model as seed points for segmenting land cover types using SAM. Fig. 2 shows some representative examples of automated segmentation with the seed points denoted by the green stars. Our preliminary results indicate that segmentations generated by SAM through this method are visually more consistent than the human-annotated polygons, especially on the forests land cover types.

As we also found that by masking the irrelevant land cover via a median mask (creating a targeted classification) can result in a significant boost in classification accuracy, increasing it from 76.91% to 82.14% on the Kahikatea dataset, this indicates that further benefits may potentially be obtained by employing an segmentation-aware approach for classification of forest cover types that employs improved automated segmentation instead of imprecise human annotations.

Currently, we are exploring an iterative annotation and training loop. We propose that the results of the automated segmentation presented in this work can be used to improve the annotations of the land cover, and the improved annotation can then be used to train the model. Ultimately, we propose that this approach would improve the quality of the classification models and annotations, and potentially reduce the need for laborious human annotations.

REFERENCES

[1] M. Pritt and G. Chern, "Satellite image classification with deep learning," in *2017 IEEE applied imagery pattern recognition workshop (AIPR)*. IEEE, 2017, pp. 1–7.

[2] E. Guirado, S. Tabik, M. L. Rivas, D. Alcaraz-Segura, and F. Herrera, "Whale counting in satellite and aerial images with deep learning," *Scientific reports*, vol. 9, no. 1, p. 14259, 2019.

[3] V. Khryashchev and R. Larionov, "Wildfire segmentation on satellite images using deep learning," in *2020 Moscow Workshop on Electronic and Networking Technologies (MWENT)*. IEEE, 2020, pp. 1–5.

[4] P. Wang, B. Bayram, and E. Sertel, "A comprehensive review on deep learning based remote sensing image super-resolution methods," *Earth-Science Reviews*, p. 104110, 2022.

[5] D. Bull, N. Lim, and E. Frank, "Perceptual improvements for super-resolution of satellite imagery," in *2021 36th International Conference on Image and Vision Computing New Zealand (IVCNZ)*. IEEE, 2021, pp. 1–6.

[6] A. A. Adegun, S. Viriri, and J.-R. Tapamo, "Review of deep learning methods for remote sensing satellite images classification: experimental survey and comparative analysis," *Journal of Big Data*, vol. 10, no. 1, p. 93, 2023.

[7] Y. Jia, E. Frank, B. Pfahringer, A. Bifet, and N. Lim, "Studying and exploiting the relationship between model accuracy and explanation quality," in *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part II 21*. Springer, 2021, pp. 699–714.

[8] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," *arXiv:2304.02643*, 2023.

[9] Q. Wu and L. P. Osco, "samgeo: A python package for segmenting geospatial data with the segment anything model (sam)," *Journal of Open Source Software*, vol. 8, no. 89, p. 5663, 2023.

[10] N. Lim, A. Bifet, D. Bull, E. Frank, Y. Jia, J. Montiel, and B. Pfahringer, "Showcasing the taiao project: providing resources for machine learning from images of new zealand's natural environment," *Journal of the Royal Society of New Zealand*, vol. 53, no. 1, pp. 69–81, 2023.

[11] "LCDB v5.0 - Land Cover Database version 5.0, Mainland, New Zealand." [Online]. Available: https://lris.scinfo.org.nz/layer/104400-lcdb-v50-land-cover-database-version-50-mainland-new-zealand/

[12] "Toitū Te Whenua Land Information New Zealand." [Online]. Available: https://basemaps.linz.govt.nz/@-41.8899962,174.0492437,z5

[13] "Land Cover Class correlations between LCDB versions." [Online]. Available: https://lris.scinfo.org.nz/layer/104400-lcdb-v50-land-cover-database-version-50-mainland-new-zealand/attachments/22492/view/

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[15] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.