Segmentation of large herbivorous mammals grazing trails

José Francisco Díez-Pastor[1][0000−0001−5013−7505], Francisco Javier González-Moya[1][0000−0002−9473−9572], Pedro Latorre-Carmona[1][0000−0001−6984−5173], Francisco Javier Pérez-Barbería[2][0000−0001−7513−5418], Ludmila Kuncheva[3][0000−0002−0415−6964], Antonio Canepa-Oneto[1][0000−0002−0608−2743], Álvar Arnaiz-González[1][0000−0001−6965−0237], and César García-Osorio[1][0000−0002−1206−1084]

[1] Department of Computer Engineering, Universidad de Burgos, Spain
jfdpastor@ubu.es
[2] Biodiversity Research Institute, Spanish Research Council, University of Oviedo, Spain
[3] School of Computer Science and Electronic Engineering, Bangor University, United Kingdom

Abstract. This is a summary of the article "Remote sensing colour image semantic segmentation of large herbivorous mammal trails" [1] published in International Journal of Remote Sensing, presented to the Multi-Conference CAEPIA'26 KeyWorks.

Keywords: Semantic segmentation · deep Learning · grazing trails · herbivory · biodiversity · monitoring

1 Introduction

Large terrestrial herbivores maintain, modify, and create new habitats and niches at various spatial scales, promoting habitat heterogeneity and, consequently, biodiversity. One of the landscape features created by the activity of large mammals, are ungulates, also known as "grazing trails". These are narrow trails shaped by continuous animal trampling in a dominant direction toward a resource or to reduce energy expenditure during motion on steep terrain.

The automatic characterisation of trails has received limited attention, with a few examples relying on landscape features such as relief, abrupt elevation changes, topographic position, among others. We improve upon this approach by using semantic segmentation, enabling pixel-level detection that ultimately provides an effective means of reconstructing the full extent of grazing trails.

2 Semantic segmentation

In this study, five semantic segmentation architectures to identify grazing trails in aerial images were used:

– U-Net (UNet) is the *seminal* work that proposed the so-called *fully convolutional network* architecture/methodology for computer vision [3].
– Feature Pyramid Networks (FPNet) are generic feature extractors that exploit multi-level feature representations in an inherent and pyramidal-type hierarchy [2].
– The Pyramid Scene Parsing Network (PSPNet) method tries to take advantage of the capability to use global context information by an aggregation methodology of different-region based features [6].
– The transformer framework for semantic segmentation (Semantic Segmentation with Transformers or Segformer) can be understood as a semantic segmentation method which unifies the so-called *transformers* with lightweight multilayer perceptron (MLP) decoders [5].
– The Unified Perceptual Parsing Network (UperNet) is a network based on FPNet [4], previously described in [2].

These methods were combined with 14 encoders, they accomplished two main purposes: (a) to generate a high-dimensional feature vector from the input image, and (b) to aggregate features across multiple resolution levels. Similarly, decoders also serve two key functions: (a) to produce a semantic segmentation mask from the high-dimensional feature vector, and (b) to decode the multi-level features aggregated by the encoder.

2.1 Experimental setup

A 10−fold cross-validation strategy was followed on every architecture-encoder pair, *i.e.,* $5 \times 14 = 70$ combinations. In every single fold, from the set of 100 aerial images (with their corresponding 100 *groundtruth* images), 80 images were used for model training, 10 images for internal model validation and, finally, 10 images were used for testing. The set of training, validation, and test images alternates on every fold to ensure the use of each instance in the test part only once. The experimentation was performed in Python using the Pytorch Segmentation Model library.

To compare the performance of the different models, the Intersection Over Union (IoU), *Precision*, *Recall*, and the *F1*-score were used.

3 Results

The proposed semantic segmentation approach effectively classified herbivore trails by accurately predicting the probability that a pixel is part of a grazing trail with high precision. Figure 1 presents examples of segmentation results produced by the optimal encoder-architecture combination (right), alongside their corresponding *groundtruth* images (left). In Figure 1 (right) the true positives are coloured in blue, false positives in red, and false negatives in yellow. A zoomed-in part has been created as well, in each one of these figures, for illustration purposes.

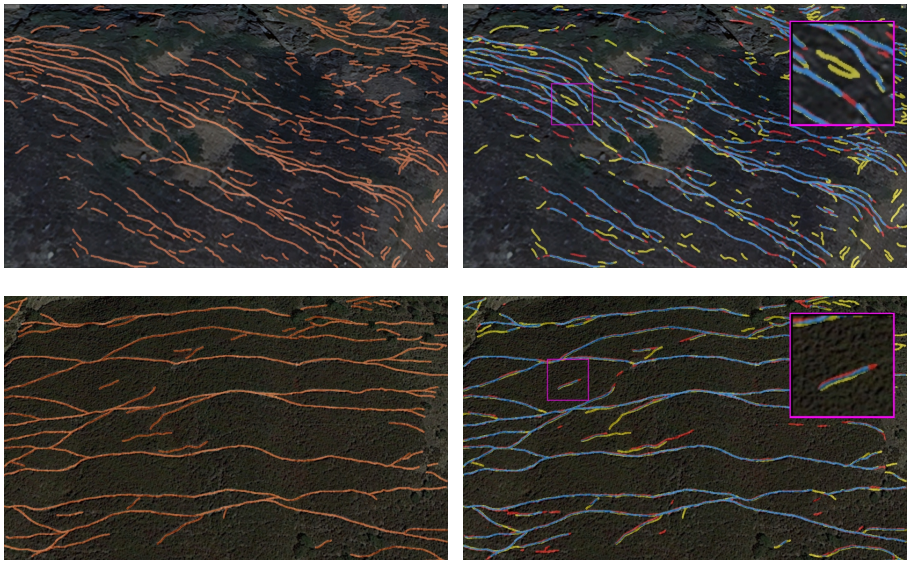


Fig. 1. Example of the original RGB *groundtruth* image (left columns) and the result of the automatic trail segmentation (right column).

Given that five architectures were tested in combination with 14 encoders, the results exhibit considerable variability. The UNet architecture with the MambaOut encoder achieved the highest overall performance across combinations of architectures and encoders, with UNet and MIT B5 encoder yielding a similar performance. Overall, the findings are consistent across the two metrics (IoU and F_1 score), with UNet achieving the best performance, and PSPNet the least. Regarding encoders, MIT B5, MambaOut, and ConvNeXt exhibit a superior performance than the rest.

4 Conclusions and future work

The study presents the first pixel-level semantic segmentation approach for mapping grazing trails created by large herbivores, advancing beyond the limitations of current patch-based detection methods.

Through the systematic evaluation of several modern semantic segmentation architectures and encoding methods, it was determined that the best model architecture is UNet, closely followed by FPNet, UperNet, and Segformer. In contrast, PSPNet showed the worst performance. Based on these findings, we recommend UNet for future studies in this domain and advise against the use of PSPNet. Additionally, another interesting result of this study is the important influence of the encoding method on model performance, with MIT B5 and MambaOut (Base) performing the best across all the architectures tested.

The work demonstrates that semantic segmentation can be effectively used to identify herbivory trails through transfer learning, even when relying on a limited set of manually labelled images.

Importantly, this study shows that accurate detection of grazing trails can be achieved using only aerial RGB imagery, without the need to incorporate ancillary information such as Digital Elevation Models or other terrain descriptors. This independence from additional geospatial layers makes the approach logistically simpler and more pragmatic. Nonetheless, future work could explicitly test the incorporation of Digital Elevation Models into the modelling framework to evaluate whether terrain slope helps to characterise distinct grazing-trail geometries that may be conditioned by topography.

Acknowledgments

This work is partially supported by Asturias Biodiversity Complementary Program BIO06 (Next Generation EU/PRTR), Strategic Projects Oriented Towards Ecological and Digital Transition (TED2021-131388B-100), Spanish Knowledge Generation Projects (PID2023-146074OB-I00) funded by EU Next Generation and Spanish Research Agency, and Spanish Research Council Tenured Scientist Incorporation Grants 2022 (202230I041).

References

1. Díez-Pastor, J.F., González-Moya, F.J., Latorre-Carmona, P., Pérez-Barbería, F.J., Kuncheva, L., Canepa-Oneto, A., Arnaiz-González, Á., García-Osorio, C.: Remote sensing colour image semantic segmentation of large herbivorous mammal trails. International Journal of Remote Sensing pp. 1–24 (2026)
2. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 936–944 (2017). https://doi.org/10.1109/CVPR.2017.106
3. Ronneberger, Olafand Fischer, P., , Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI. pp. 234–241. Springer Berlin Heidelberg, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
4. 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). pp. 432–448. Springer International Publishing, Cham (September 2018)
5. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 12077–12090. Curran Associates, Inc. (2021)
6. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6230–6239. IEEE Computer Society, Los Alamitos, CA, USA (jul 2017). https://doi.org/10.1109/CVPR.2017.660