Testing the transferability of CarcassonNet

A case study to detect hollow roads in Germany and Slovenia

Wouter VERSCHOOF-VAN DER VAART, Leiden University, Faculty of Archaeology, The Netherlands
Juergen LANDAUER, Landauer Research, Germany

Abstract: While (post)medieval hollow roads (sunken cart tracks) are hard to discern in the present-day landscape, they appear as distinct lines in LiDAR data. This offers opportunities to use automated detection methods to map these archaeological objects. However, the usability of such methods when used on an area unrelated to the area in which the method was developed, i.e., the transferability or generalization capability, remains an unanswered question. Therefore, in this paper, the transferability of CarcassonNet, a workflow combining a Deep Learning Convolutional Neural Network and image processing algorithms to map hollow roads in LiDAR data, is tested. CarcassonNet is trained on LiDAR data from the Veluwe region in the central part of the Netherlands. Subsequently, the workflow is tested on LiDAR data deriving from the Schurwald region in southern Germany and the Upper Carniola and Nova Gorica regions in northern and southwestern Slovenia respectively. These areas have different terrain and land-use compared to the Veluwe while the properties of the LiDAR data also differ. The results show that CarcassonNet is able to detect hollow roads as long as threshold moving is applied and the confidence threshold is re-determined. Differences in terrain and land-use seem to be of minor influence on the performance of CarcassonNet. However, it is apparent that differences in the quality of the LiDAR data, most probably the difference in average ground point density, do influence performance.

Keywords: *LiDAR—CNN—Hollow roads—Machine Learning—Transferability*

CHNT Reference: Verschoof van der Varart, W. and Landauer, J. (2022). 'Testing the transferability of CarcassonNet. A case study to detect hollow roads in Germany and Slovenia', in Börner, W., Rohland, H., Kral-Börner, C. and Karner, L. (eds.) *Proceedings of the 25th International Conference on Cultural Heritage and New Technologies, held online, November 2020*. Heidelberg: Propylaeum. doi:10.11588/propylaeum.1045.c14482

Introduction

Medieval hollow roads are linear, sunken trails caused by continuous passage of carts and other traffic through the landscape (Kirchner et al., 2020). These tracks are hard to discern in the present-day terrain but appear as distinct longitudinal objects in airborne Light Detection and Ranging (LiDAR) data (Kokalj and Hesse, 2017; Figure 1). This marked appearance offers opportunities to use Deep Learning methods to systematically map these hollow roads and reconstruct the historical route network.

The results of such endeavours can supplement and expand the knowledge gained from historical written sources and cartographic data (Kirchner et al., 2020). In archaeology, Deep Learning has successfully been implemented to detect discrete objects, such as barrows (Verschoof-van der Vaart et al., 2020), but has been scarcely used for more complex, large-scale patterns, such as roads (Traviglia and Torsello, 2017; Davis, 2021).

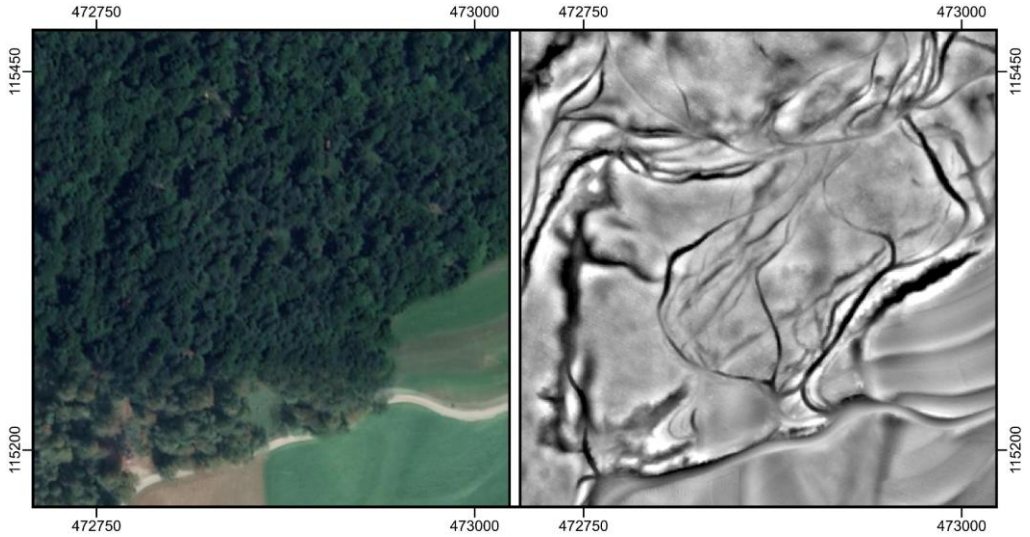

Fig. 1. An example of hollow roads on a recent aerial photograph (left) and on LiDAR data (right), visualized with SLRM (Hesse, 2010), from the Upper Carniola region, Slovenia (Coordinates in MGI 1901/Slovene National Grid, EPSG: 3912; © Authors).

This is in part because modern road detection methods (Abdollahi et al., 2020) do not translate well to the problem of hollow roads—typically several linear tracks with slightly different orientation and multiple overlaps—that lack the uniformity in shape and appearance of modern roads. Furthermore, hollow roads are often only partially preserved and regularly are dissected by modern landscape objects (Verschoof-van der Vaart and Landauer, 2021).

Therefore, to map hollow roads in LiDAR data, a workflow that combines Deep Learning and image processing algorithms has been developed (Verschoof-van der Vaart and Landauer, 2021). This workflow, named CarcassonNet, has successfully been trained and evaluated on the *Veluwe* area in the central part of the Netherlands. However, the transferability of CarcassonNet (or automated detection methods in general), i.e., the usability of the method on an unrelated area with either different topography, land-use and/or LiDAR data of different properties, remains an unanswered question in this research field (Kermit, Reksten and Trier, 2018; Cowley et al., 2020). As can be imagined, these factors can vary considerably on a regional or national or even international level, and a method that is unable to adjust to such different situations has little practical value for wide application. Therefore, studies in different environments are important to investigate the true potential of automated approaches in archaeological practice. In this paper, the transferability of CarcassonNet is investigated, i.e., whether our method, trained on Dutch LiDAR data, is able to detect hollow roads in LiDAR data from Germany and Slovenia.[1]

Research areas

CarcassonNet has been trained on LiDAR data from the western part of the province of Gelderland in the Netherlands, known locally as the *Veluwe* (see Table 1). This region consists of several north-south orientated ice-pushed ridges separated by relatively flat valleys. The area used (circa 97 km^2) is predominantly covered with heath and to a lesser extent with forest, and is interspersed with agricultural fields and areas of habitation (for a detailed overview of the Veluwe, see Lambers,

[1] We are grateful to R. Hesse (Landesamt für Denkmalpflege Baden-Württemberg) and Ž. Kokalj (Research Centre of the Slovenian Academy of Sciences and Arts) for providing LiDAR data to test CarcassonNet.

Verschoof-van der Vaart and Bourgeois, 2019). The model has been tested on LiDAR data from the *Schurwald* region (Germany) and the *Upper Carniola* and *Nova Gorica* regions (Slovenia; see Table 1). The Schurwald is a wooded mountain range in the federal state of Baden-Württemberg in southern Germany. The area used (25 km^2) is covered with forest, crossed by numerous eroded stream valleys, and has various areas that have been cleared for agriculture and occupation (Hesse, 2013). The Upper Carniola region is a rugged area in the foothills of the Kamnik–Savinja Alps in northern Slovenia. The area used (circa 1.2 km^2) is predominantly covered with forest, villages, and some agricultural fields (Kokalj and Hesse, 2017). The Nova Gorica region lies on the border of the Karst Plateau in southwestern Slovenia. High lying plateaus with steep declines and many heavily eroded stream valleys characterize this landscape. The area used (circa 1.3 km^2) is covered with forest. In part of the area military trenches, comparable in morphology to hollow roads, are present (Kokalj and Hesse, 2017). All hollow roads in the Schurwald, Upper Carniola, and Nova Gorica area have been manually annotated by the authors, who both have ample experience in analysing LiDAR data.

The LiDAR data from the Veluwe, used for the training of CarcassonNet, has a raster resolution of 0.5 m and an average ground point density of 6–10 points per square meter. The resolution of the test data differs between 0.5 m (Upper Carniola) and 1.0 m (Schurwald and Nova Gorica). The average ground point density of the Slovenian data varies between 3 and 10 points per square meter, while the data from the Schurwald has a lower point density between 1 and 4 points per square meter (Table 1).

Table 1. The LiDAR datasets used in this research to train and test CarcassonNet.

Dataset	Area (sq. km)	Resolution	Average Point-density (pt. per sq. m)	Mean elevation	Min–Max elevation	General terrain	Main land-use
Veluwe (NL)	97.25	0.5	6–10	42	10–104	Ridges/valleys	Heath
Schurwald (DE)	25	1.0	1–4	172	13–254	Mountains	Forest
Upper Carniola (SI)	1.18	0.5	3–10	393	342–487	Hills	Forest
Nova Gorica (SI)	1.32	1.0	3–10	115	67–190	Karst	Forest

Methodology

The CarcassonNet workflow (Fig. 2) consists of three steps: preprocessing, classification, and post-processing (for an extensive overview of the methodology of CarcassonNet see (Verschoof-van der Vaart and Landauer, 2021). An important innovation in the CarcassonNet workflow is that multiple sections per single hollow road, as opposed to the entire road, are used as input images (comparable to the approach taken for Celtic fields in Verschoof-van der Vaart et al., 2020). This makes it much more cost-effective to create a sufficient training dataset.

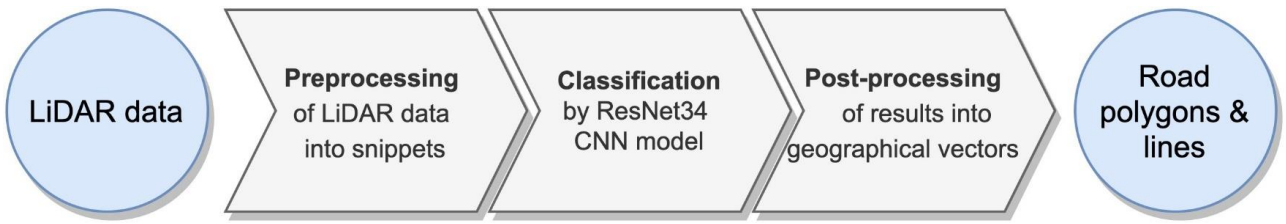

Fig. 2. Simplified representation of the CarcassonNet workflow; after Verschoof-van der Vaart and Landauer, 2021.

To create a training dataset, the first author manually annotated the LiDAR data from the Veluwe (97.25 km^2) in order to create a reference standard. Subsequently, in the preprocessing step the same LiDAR data was split into snippets of 64 by 64 pixels. For every snippet the percentage of overlap with the roads polygons in the reference standard was computed. Snippets with an overlap of 95% or more with the mapped hollow roads were put in a 'hollow road' class and only those with an overlap of 5% or less into the 'empty' class to create a binary dataset. Additional measures to balance the dataset resulted in training dataset of circa 32,000 snippets, with circa 16,000 snippets in both classes.

The second step of CarcassonNet consists of a Resnet-34 Convolutional Neural Network (He et al., 2016). A CNN is a hierarchically structured (image) feature extractor and classifier algorithm, loosely inspired by the animal visual cortex (Guo et al., 2016). These algorithms learn to generalize from given examples, i.e., a large set of labelled images, rather than relying on a human operator to set parameters or formulate rules. Furthermore, CNNs can be pre-trained on a large, generic dataset (for instance images of cats and dogs) and subsequently transfer-learned on a small, specific dataset (such as our dataset). In this research the CNN was only used for binary classification, a relatively simple task, which produces better detection results at a lower cost and effort (Guo et al., 2016), as opposed to a more complicated task such as segmentation. The ResNet-34 CNN was pre-trained on the ImageNet dataset (Russakovsky et al., 2015) and subsequently transfer-learned for twelve epochs on the LiDAR data from the Veluwe, following the "progressive resizing" training scheme recommended by the research institute *Fast.ai*. After training the CNN was tested on the data from Germany and Slovenia. To make the output of the CNN directly usable in a GIS environment, the detections are turned into geospatial vectors (polygons and lines; see Figure 3) by combining geo-spatial processing tools in *QGIS 3.4* and the image processing approach taken by Van Etten (2019).

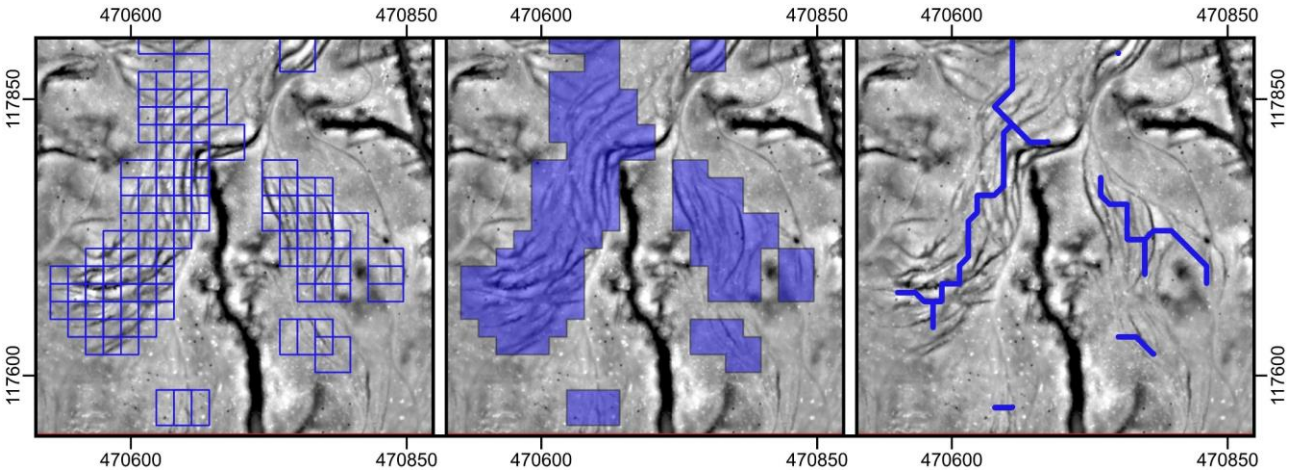

Fig. 3. Excerpts of LiDAR data, visualized with SLRM (Hesse, 2010), from the Upper Carniola region (Slovenia), showing: the results of the classification (left); the derived polygons indicating the location of the hollow roads (centre); and the derived lines depicting the route network (right; coordinates in MGI 1901/Slovene National Grid, EPSG: 3912; © Authors).

Results

To evaluate CarcassonNet, the area (in m^2) of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) were determined following the approach taken in Verschoof-van der Vaart et al., 2020. Subsequently, Matthews Correlation Coefficient (MCC; Eq. 1) was calculated.

This metric is a reliable measure of the correlation between the observed and predicted binary classification, even if the classes are very imbalanced (Luque et al., 2019). Therefore, MCC is a better indicator of the quality of CarcassonNet, compared to other metrics such as F1-score or accuracy (Chicco and Jurman, 2020). MCC is bound between -1 and 1, where higher values indicate a better performance. The default confidence threshold for classification is typically set to 0.5. However, by changing the threshold (called threshold moving) and recalculating the performance metric, an optimal trade-off can be found, resulting in the highest MCC (Zou et al., 2016). During this research the optimal confidence threshold was empirically calculated (see Table 2). Finally, for comparison of the performance between the different datasets and other methods, the well-known metrics Precision (Eq. 2) and Recall (Eq. 3) were calculated as well (but see Luque et al., 2019). Precision measures how many of the selected items are relevant. Recall gives a measure of how many relevant objects are selected. The results of the experiments are shown in Table 2 and Figure 4. CarcassonNet has moderate to high performance on the Upper Carniola and Nova Gorica datasets, reaching a MCC of 0.37 (confidence threshold of 0.1) and 0.38 (confidence threshold of 0.8) respectively. However, the performance on the Schurwald dataset is lower, reaching a MCC of 0.25 with a confidence threshold of 0.75.

Equation 1:

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

Equation 2:

$$Precision = \frac{TP}{(TP + FP)}$$

Equation 3:

$$Recall = \frac{TP}{(TP + FN)}$$

Table 2. Results of the experiments on the different datasets.

Testset	Confidence	TP	FP	TN	FN	Recall	Precision	MCC
Veluwe (NL)[2]	0.8	4,967,470	7,243,900	88,839,400	2,664,868	0.65	0.41	0.47
Schurwald (DE)	0.75	320,430	243,300	22,486,100	1,950,250	0.14	0.57	0.25
Upper Carniola (SI)	0.8	79,500	40,823	732,832	146,928	0.35	0.66	0.38
Nova Gorica (SI)	0.1	88,280	81,928	884,790	120,283	0.42	0.51	0.37

Discussion

The results of the generalization experiments (Table 2) show that CarcassonNet, when trained on LiDAR data from the Netherlands, is able to detect hollow roads in other areas with different terrain, land-use, and on LiDAR data with different properties. However, the performance of the model decreases by 9–10 points on both Slovenian datasets, while the performance on the Schurwald dataset is 22 points lower. An important note to make is that the areas from Slovenia used for testing (2.5 km² in total) are considerably smaller than the Schurwald area (25 km²). This size difference might have

[2] For comparison,the results of testing on the Veluwe dataset are reproduced from Verschoof-van der Vaart and Landauer (2021).

been of influence on the results, as the performance of a detection method can vary significantly between test datasets that have a different density of archaeological objects or in which the state of preservation of these objects varies (see Verschoof-van der Vaart et al., 2020).

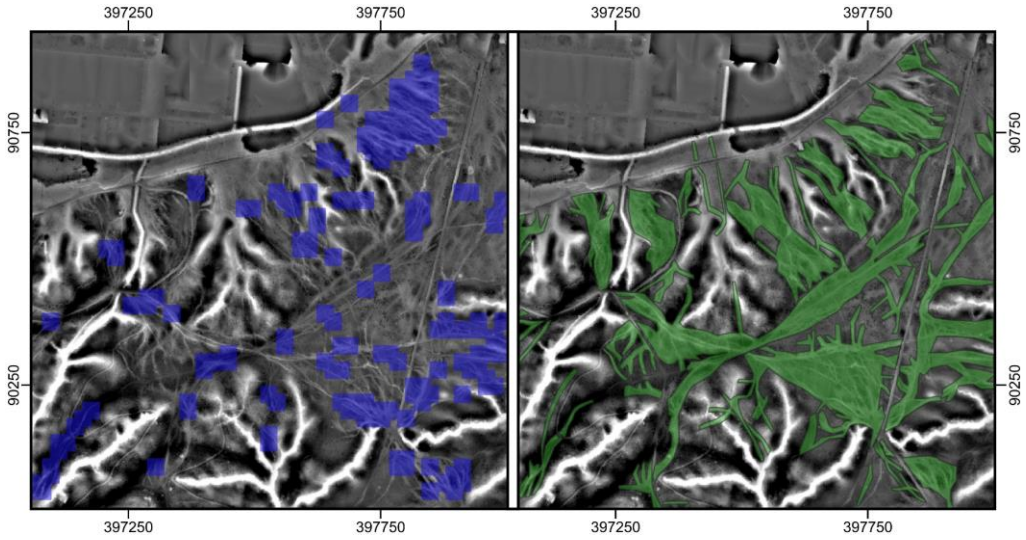

Fig. 4. Excerpts of LiDAR data, visualized with SLRM (Hesse, 2010), from the Nova Gorica region (Slovenia), showing: the results of the classification (left) and the manually annotated roads (right); coordinates in MGI 1901/Slovene National Grid, EPSG: 3912; © Authors).

The influence of the different terrain and land-use seems to have been minor, as no increase in FP is observed—the Precision is comparable for the German and Slovenian datasets and even better than on the Veluwe dataset. Therefore, a different cause needs to be sought for the lower performance, especially on the Schurwald data. Table 2 shows that the Recall on the Slovenian and in particular the Schurwald datasets is much lower, indicating that the recognition of hollow roads is the main detrimental factor. It was expected that this was caused by the difference in raster resolution between the datasets, which varies with a factor of two (0.5 versus 1.0 m). However, the comparable performance on the Nova Gorica (1.0 m resolution) and Upper Carniola (0.5 m resolution) datasets seems to indicate that this is also not of (major) influence. However, a comparison between the appearance of hollow roads in the different datasets shows that these occur much less pronounced in the Schurwald dataset (Fig. 5), which is probably related to the difference in average ground point density. The influence of ground point density on the ability of automated methods and humans to detect archaeological objects in LiDAR data has also been observed in other research, both for automated detection methods (Trier and Pilø, 2012; Dolejš et al., 2020) as well as for humans (Risbøl et al., 2013). Therefore, differences in LiDAR parameters, especially the ground point density, are probably detrimental on the performance of CarcassonNet.

Conclusion

This research shows that CarcassonNet, when trained on LiDAR data from the Netherlands, is able to generalize and detect hollow roads in data from Germany and Slovenia even though the areas have different terrain, land-use, and the LiDAR data has different properties. However, the performance of the model does decrease, in the case of the Schurwald area to the point that it is questionable if the method would be usable. Probably the difference in the average ground point density of the LiDAR datasets is the main negative influence on the performance. Therefore, further research

will focus on improving CarcassonNet, to better cope with hollow roads in LiDAR data with different properties. For instance, by combining LiDAR datasets with varying properties and/or from different areas. Also the potential of Generative Adversarial Networks to increase the image resolution, and therefore increase the visibility of hollow roads in the LiDAR data will be explored (Ledig et al., 2017).

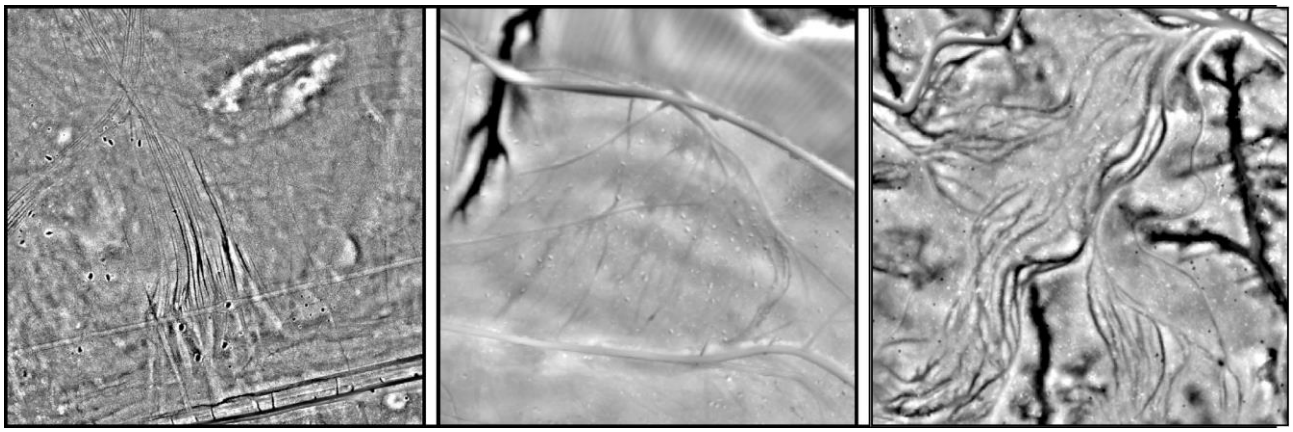

Fig. 5. Excerpt of LiDAR data, visualized with SLRM (Hesse, 2010), showing the difference in appearance of hollow roads in the different areas, from left to right: Veluwe, Schurwald, Upper Carniola (scale 1:5000; © Authors).

Funding

This research was in part supported by the Data Science Research Programme (Leiden University).

Conflict of Interests Disclosure

The authors have no competing interests to declare.

Author Contributions

Conceptualization, Data curation, Formal Analysis. Investigation, Methodology, Project Administration, Resources, Supervision, Validation, Writing – original draft, Writing – review & editing: WV & JL
Visualization, Funding acquisition, Project Administration: WV
Software: JL

References

Abdollahi, A. et al. (2020) 'Deep Learning Approaches Applied to Remote Sensing Datasets for Road Extraction: A State-Of-The-Art Review', *Remote Sensing*. Multidisciplinary Digital Publishing Institute, 12(9), p. 1444. doi:10.3390/rs12091444.

Chicco, D. and Jurman, G. (2020). 'The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation', *BMC Genomics*, 21(6), pp. 1–13. doi:10.1186/s12864-019-6413-7.

Cowley, D. et al. (2020). 'Making LiGHT Work of Large Area Survey? Developing Approaches to Rapid Archaeological Mapping and the Creation of Systematic National-scaled Heritage Data', *Journal of Computer Applications in Archaeology*, 3(1), pp. 109–121. doi:10.5334/jcaa.49.

Davis, D. S. (2021). 'Theoretical Repositioning of Automated Remote Sensing Archaeology: Shifting from Features to Ephemeral Landscapes', *Journal of Computer Applications in Archaeology*, 4(1), pp. 94–109. doi:10.5334/jcaa.72.

Dolejš, M. et al. (2020). 'Aerial Bombing Crater Identification: Exploitation of Precise Digital Terrain Models', *ISPRS International Journal of Geo-Information*, 9(12), p. 713. doi:10.3390/ijgi9120713.

Van Etten, A. (2019). 'City-scale Road Extraction from Satellite Imagery'. Available at: http://arxiv.org/abs/1904.09901.

Guo, Y. et al. (2016). 'Deep learning for visual understanding: A review', *Neurocomputing*, 187, pp. 27–48. doi:10.1016/j.neucom.2015.09.116.

He, K. et al. (2016). 'Deep Residual Learning for Image Recognition', in *2016 IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV: IEEE, pp. 770–778. doi:10.1109/CVPR.2016.90.

Hesse, R. (2010). 'LiDAR-derived local relief models-a new tool for archaeological prospection', *Archaeological Prospection*, 17(2), pp. 67–72. doi:10.1002/arp.374.

Hesse, R. (2013). 'The changing picture of archaeological landscapes: Lidar prospection over very large areas as part of a cultural heritage strategy', in Opitz, R. and Cowley, D. (red.) *Interpreting Archaeological Topography: 3D data, Visualisation and Observation*. Oxford: Oxbow, pp. 171–183.

Kermit, M., Reksten, J. H. and Trier, Ø.D. (2018). 'Towards a national infrastructure for semi-automatic mapping of cultural heritage in Norway', in Matsumoto, M. and Uleberg, E. (red.) *Oceans of Data. Proceedings of the 44th Conference on Computer Applications and Quantitative Methods in Archaeology*. Oxford: Archaeopress, pp. 159–172.

Kirchner, A. et al. (2020). 'Spatial analysis of hollow ways in the Hildesheimer Wald Mountains (Lower Saxony, Germany) as a model for mountainous regions of Central Europe', *Erdkunde*, 74(1), pp. 1–14. doi:10.3112/erdkunde.2020.01.01.

Kokalj, Žiga and Hesse, R. (2017). *Airborne Laser Scanning Raster Data Visualization: A Guide to Good Practice*. Ljubljana: Založba ZRC.

Lambers, K., Verschoof-van der Vaart, W. B. and Bourgeois, Q. P. J. (2019). 'Integrating remote sensing, machine learning, and citizen science in Dutch archaeological prospection', *Remote Sensing*, 11(7), p. 794. doi:10.3390/rs11070794.

Ledig, C. et al. (2017). 'Photo-realistic single image super-resolution using a generative adversarial network', in *Proceedings – 30th IEEE Conference on Computer Vision and Pattern Recognition*, pp. 105–114. doi:10.1109/CVPR.2017.19.

Luque, A. et al. (2019). 'The impact of class imbalance in classification performance metrics based on the binary confusion matrix', *Pattern Recognition*. Elsevier Ltd, 91, pp. 216–231. doi:10.1016/j.patcog.2019.02.023.

Risbøl, O. et al. (2013). 'Interpreting cultural remains in airborne laser scanning generated digital terrain models: effects of size and shape on detection success rates', *Journal of Archaeological Science*, 40(12), pp. 4688–4700. doi:10.1016/j.jas.2013.07.002.

Russakovsky, O. et al. (2015). 'ImageNet Large Scale Visual Recognition Challenge', *International Journal of Computer Vision*. Springer New York LLC, 115(3), pp. 211–252. doi:10.1007/s11263-015-0816-y.

Traviglia, A. and Torsello, A. (2017). 'Landscape Pattern Detection in Archaeological Remote Sensing', *Geosciences*, 7(4), p. 128. doi:10.3390/geosciences7040128.

Trier, Øivind Due and Pilø, L. H. (2012). 'Automatic Detection of Pit Structures in Airborne Laser Scanning Data', *Archaeological Prospection*, 19(2), pp. 103–121. doi:10.1002/arp.1421.

Verschoof-van der Vaart, W. B. et al. (2020). 'Combining Deep Learning and Location-Based Ranking for Large-Scale Archaeological Prospection of LiDAR Data from The Netherlands', *ISPRS International Journal of Geo-Information*, 9(5), p. 293. doi:10.3390/ijgi9050293.

Verschoof-van der Vaart, W. B. and Landauer, J. (2021). 'Using CarcassonNet to automatically detect and trace hollow roads in LiDAR data from the Netherlands', *Journal of Cultural Heritage*, 47, pp. 143–154. doi:10.1016/j.culher.2020.10.009.

Zou, Q. et al. (2016). 'Finding the Best Classification Threshold in Imbalanced Classification', *Big Data Research*, 5, pp. 2–8. doi:10.1016/j.bdr.2015.12.001.