Learning to look at LiDAR: combining CNN-based object detection and GIS for archaeological prospection in remotely-sensed data

Verschoof-van der Vaart, W.B.

Part III

Discussion

7

Discussion

> "The trouble with computers, of course, is
> that they're very sophisticated idiots."
>
> **Doctor Who**
> (Doctor Who, Season 12,
> Episode 4: "Robot: Part Four")

7.1 Introduction

The importance of remotely-sensed data, especially LiDAR data, for archaeology has grown exponentially in recent years (Opitz & Herrmann, 2018). Nowadays, the manual analysis of these data sources is a key element of local and regional scale archaeological research, as well as cultural heritage management (Cowley & Sigurdardóttir, 2011; Verhoeven, 2017). However, to overcome the challenges to manually analyze big remotely-sensed datasets and to find and document the seemingly overwhelming number of potential archaeological objects (Bennett et al., 2014; Bevan, 2015), advanced computational or manual brute-force search strategies are needed (Casana, 2014, 2020).

Therefore, archaeologists have started developing computational methods for the (semi-)automated detection of archaeological objects (Lambers et al., 2019). Since then multiple studies have shown that these algorithms are capable of detecting well-defined archaeological objects (see Chapter 1, Fig. 1.2 for an overview). However, these (often) handcrafted algorithms are highly specialized on specific, single object categories and data sources, which restricts their use in different contexts and limits their usability in general for archaeological prospection. The recent developments in Deep Learning (LeCun et al., 2015), instigated by the emergence of Convolutional Neural Networks (CNNs; Krizhevsky et al., 2012), have shown the potential of these methods in multiple domains, for a variety of tasks (Perrault et al., 2019), including remote-sensing (Ball et al., 2017) and archaeology (Trier et al., 2018; Zingman, 2016). However, many challenges remain concerning the application of Deep Learning approaches in 'real-world' scenarios, including archaeological practice. Therefore, the first aim of this thesis has been to develop and apply methodologies using Deep (Region-based) CNNs for the detecting of (multiple classes of) archaeological objects in LiDAR data. The second aim was to explore the usability and incorporation of these CNN-based methodologies in archaeological practice and spatial archaeology. The ultimate goal of this research is to advance the use of Deep Learning and LiDAR in archaeology.

This chapter starts with a summary of the challenges of archaeological object detection in remotely-sensed data and the progress made towards resolving these (Section 7.2). One of the main outcomes of this research, a technique to add domain knowledge to the classification process, is discussed in more detail (Section 7.2.3). Subsequently, in Section 7.3 the evaluation of the performance of these methods and their transferability will be reviewed. Afterwards, the incorporation of these methods in archaeological practice is considered (Section 7.4). The chapter concludes with a summary of the most important insights generated by the research presented in this thesis (Section 7.5) and suggestions for future work (Section 7.6).

7.2 Archaeological Object Detection in Remotely-Sensed Data

The first aim of this thesis has been to investigate the possibilities and challenges of using Deep (Region-based) CNNs for archaeological object detection in remotely-sensed data. However, this particular task is not as straightforward as more general object detection tasks, such as finding persons, animals, or household objects in photographs (see Everingham et al., 2010; Lin et al., 2014b). Consequently, the implementation of traditional object detection methods is limited by: 1) the nature and appearance of archaeological objects in remotely-sensed data; 2) their similarity to other anthropogenic and natural landscape elements; and 3) the characteristics of the remotely-sensed data and images. These challenges, which are not restricted to archaeology but are also prevalent in other domains (Sumbul et al., 2019; Tang et al., 2017a; Van Etten, 2018), will be discussed in more detail, followed by an overview of the solutions and workflows developed in this research.

7.2.1 Challenges

Small and Scarce Objects

The main challenge in using traditional object detection methods for our particular task is that archaeological objects in remotely-sensed images are usually small, generally lack a consistent orientation, and are often densely clustered but scarcely distributed. Furthermore, their degree of preservation varies greatly and with it, their appearance in these images. For example, in our random test dataset from the Veluwe the size of barrows and charcoal kilns is on average only 14–20 m (26–40 pixels), while plots within Celtic fields are slightly larger with sizes of circa 56 m (112 pixels; see Fig. 7.1). In comparison, the coverage of a single image in the random test dataset is 300 by 300 m (600 by 600 pixels) or 90,000 m^2! While the Veluwe area holds one of the densest concentrations of archaeological objects in the Netherlands, these objects nevertheless only appear in 1 in every 5 images in the random test dataset (see Chapter 3, Table 3.2). Detecting objects in such a 'low density' dataset is a challenging task, not only for automated detection methods but also for human interpreters (Soroush et al., 2020).

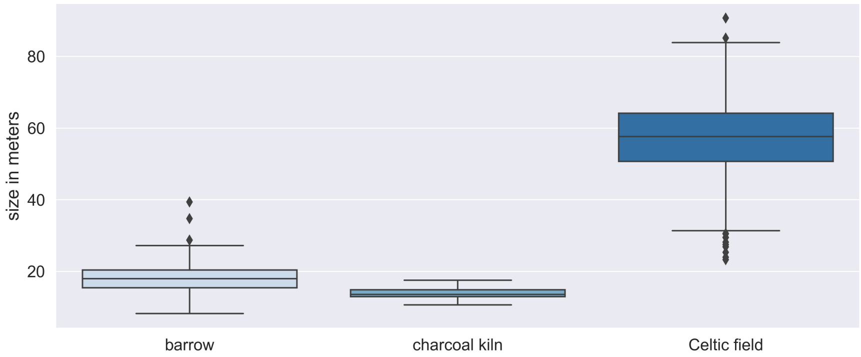

Figure 7.1: Boxplots showing the average size of objects in the random test dataset.

This is in stark contrast to more general purpose datasets used for CNN-based object detection, such as Microsoft COCO (Lin et al., 2014b) or Pascal VOC (Everingham et al., 2010), which contain 'natural images', i.e., photographs of scenes seen in normal settings (Goodfellow et al., 2016). These images normally contain large, prominent objects (i.e., they occupy a major portion of the image) that are overall reliably orientated (Fig. 7.2). Traditional object detection methods take advantage of these relatively large objects by heavily downscaling the images when they pass through the CNN, which greatly reduces the computational cost (Olivier & Verschoof-van der Vaart, 2021). Therefore, directly applying CNN-based object detection algorithms renders unsatisfactory performance when applied to detect small objects in images, as these methods are geared towards detecting abundantly present, large objects (see Everingham et al., 2015; Ren et al., 2018b). Furthermore, small and scarcely distributed objects lead to the problem of foreground-background class imbalance in object detection (Oksuz et al., 2019), where one class is over-represented, in this case the background class, while the other class (foreground, i.e., the archaeological objects) is under-represented (Luque et al., 2019). This imbalance can have a major impact on the classification and generalization capacity of CNNs, leading to bias and low performance (see also Section 7.3.2).

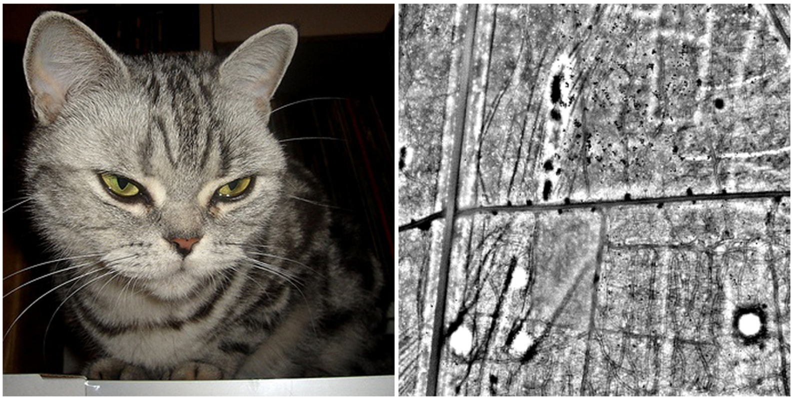

Figure 7.2: A comparison of a natural image from the Pascal VOC 2007 dataset (left; Everingham et al., 2010) versus an extract of LiDAR data of the same pixel size, visualized with Simple Local Relief Model (Hesse, 2010), showing barrows, Celtic fields, and hollow roads (source of the height model: Nationaal Georegister, 2021).

Landscape Patterns

Automated detection methods have mainly been developed to find compact, localized, and discrete objects, such as persons and household objects, or barrows and charcoal kilns in the case of archaeology (Davis, 2021). However, in archaeological spatial analysis more complex, large-scale landscape patterns, such as field systems and roads, are also discerned and documented (Doneus, 2013; Traviglia & Torsello, 2017). These landscape patterns generally consist of a system of related entities (e.g., plots within Celtic fields) as opposed to separate, single entities.

From a more theoretical point of view, these objects transcend from being delimited points of interest in the landscape to being the areas or paths between these points, but are of no less interest than the places they surround or connect (see Chapman, 2011; Doneus, 2013). These objects pose a challenge for traditional object detection methods as these traces generally cover extensive areas and/or have irregular shapes, which makes it problematic to 'catch' these in (rectangular) bounding boxes. In that sense the problem of detecting landscape patterns is more closely related to land cover classification (Vali et al., 2020). A possible solution to this problem is to use rotatable bounding boxes (Yu et al., 2020), although this does not solve the problem of long linear traces, e.g., roads. Furthermore, these patterns are often difficult to define (on a pixel level), due to their heterogeneous nature (see Guyot et al., 2018). Finally, these objects generally lack the uniformity of their modern counterparts, are often only partially preserved due to anthropogenic and geomorphological interference, and are regularly dissected by modern landscape objects (see Fig. 7.2). Therefore, approaches developed to directly detect modern versions of these landscape patterns, for instance motorways, do not translate well to the problem at hand (see Chapter 5). Alternatively, in archaeology different methods have been developed that use specialized (hand-crafted) feature extraction and analysis techniques (Figorito & Tarantino, 2014; Kirchner et al., 2020; Traviglia & Torsello, 2017; Vletter, 2014; Vletter & van Lanen, 2018) or that indirectly detect landscape patterns using spatial and predictive modeling (Davis et al., 2020; Verhagen et al., 2019) or ontological reasoning (Nuninger et al., 2020a,b).

Objects of Confusion

Next to the problems related to the appearance of archaeological objects, a recurrent issue are anthropogenic or natural landscape elements with a comparable morphology to the archaeological objects of interest, which cause False Positives (Casana, 2020). Examples of these 'objects of confusion' are rounded outcrops or roundabouts classified as barrows (Cerrillo-Cuenca, 2017; Chapter 2), spoil heaps classified as charcoal kilns (Schneider et al., 2015), and forest planting ditches classified as hollow roads (Chapter 5.3.2). The occurrence of these objects can lead to reduced performance in automated detection methods, especially when the archaeological classes of interest are relatively 'simple' objects, e.g., mounds or pits, of which the shape appears abundantly in the landscape (Meyer-Heß, 2020). This issue increases with scale and becomes especially prevalent when methods are applied on large areas with different types of complex terrain (see Trier et al., 2021; Chapter 3.6). Contrary, human interpreters can usually differentiate between objects of confusion and archaeological objects with minor difficulty. For instance, humans seldom classify a roundabout as a barrow, even though these objects appear the same in LiDAR data, i.e., as a circular, positive elevation (Fig. 7.3). This can partly be explained by the fact that human interpreters can observe the vicinity of potential objects, i.e., they have contextual information (see also Wu et al., 2020), and that during manual analysis regularly additional data sources are consulted to make an informed decision.

In contrast, most automated detection approaches classify small segments of an image, derived from a single data source (e.g., LiDAR data), which limits the available contextual information. Furthermore, an interpreter has a certain amount of prior experience, archaeological and geological knowledge, and a degree of flexible and creative reasoning that can be used in the classification task (Cowley, 2012; White, 2019, Chapters 3.6 & 4.4). In that sense a human uses a combination of perception and comprehension to analyze the data (Lozic & Štular, 2021), while an automated detection approach only uses perception, but generally lacks comprehension. It is therefore hardly surprising that these methods struggle with objects of confusion.

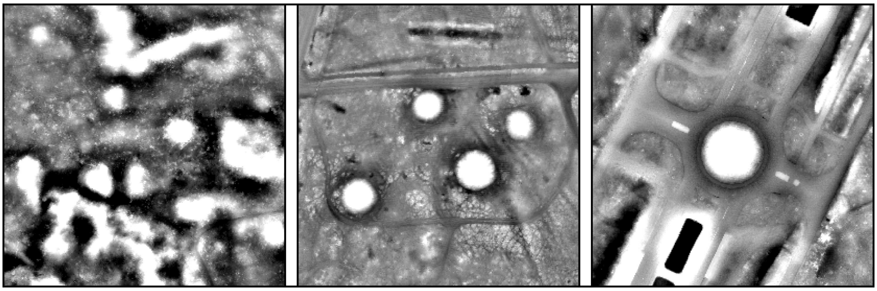

Figure 7.3: Extracts of LiDAR data, visualized with Simple Local Relief Model (Hesse, 2010), showing drift-sand dunes (left), barrows (middle), and a roundabout (right; source of the height model: Nationaal Georegister, 2021).

Remotely-Sensed Datasets versus General Purpose Datasets

Next to these challenges related to the appearance and nature of archaeological objects, there is also a difference between remotely-sensed images and the natural images normally used to train and test traditional object detection methods. Remotely-sensed images are on average much larger than natural images and frequently exceed the maximum input sizes of CNNs. For example, the LiDAR data used in this research is distributed in images of 10,000 by 12,500 pixels. Yet, the maximum input size of Faster R-CNN is 600 pixels. Therefore, images need to be split into smaller parts (i.e., subtiles or snippets), preferably with overlap. This process can result in the dissecting of archaeological objects on the edge of subtiles (see Chapter 2.4). Also, this cutting up can lead to class imbalance between the number of 'positive' and 'negative' examples due to the scarcely distributed objects.

More specifically, LiDAR images differ from natural images (RGB-colored photographs), as the former are the product of the conversion of raster datasets, containing elevation values, into human-readable grayscale (or color) images through different filtering, interpolation, and visualization techniques (Kokalj & Hesse, 2017; Opitz, 2013). Rather than having objects standing opaquely before a background, LiDAR images contain intricate surface textures that semi-transparently overlay the natural background and are in turn (partially) overlapped by other textures (Mlekuž, 2013a; see Fig. 7.2). The latter phenomenon, occlusion—when objects are not perfectly visible and are obscured by objects or vegetation—is a common issue in remotely-sensed images and problematic for detection methods (Ren et al., 2018a).

The difference between grayscale (single-channel data) and RGB-color (three-channel data) is normally resolved by turning the grayscale images into RGB by copying the value from the first channel to the other two color channels (Chollet, 2015). Although this is regarded as a valid solution, the impact on the performance of a transfer-learned CNN, pre-trained on RGB-colored images, remains a point of discussion (Xie & Richmond, 2019). Because the image characteristics of both types of imagery, grayscale LiDAR versus RGB-colored photographs, are quite different, it has been suggested that the effectiveness of transfer-learning declines as primary data, i.e., the images used for pre-training, and secondary data, i.e., the images used for fine-tuning, become less similar (Pires de Lima & Marfurt, 2019). Therefore, there might be potential for improvement in transfer-learning by using CNNs that are pre-trained on images more similar to the used remotely-sensed data (Gallwey et al., 2019; Opitz & Herrmann, 2018; Trier et al., 2021). However, to date no large remotely-sensed datasets with annotated archaeological objects (Opitz & Herrmann, 2018) and few labeled, more general remotely-sensed datasets are publicly available for pre-training (Sumbul et al., 2019). Besides, the question remains—not only in archaeology but also in Deep Learning research in general—whether pre-training on more comparable data actually improves performance (Ball et al., 2017; He et al., 2019; Trier et al., 2019; Zoph et al., 2020). In the research of Gallwey et al. (2019) it was shown that a CNN named DeepMoon (Silburt et al., 2019), originally trained to detect craters in lunar LiDAR data, reached high performance when transfer-learned to detect mining pits in terran LiDAR data. However, it remains the question whether the pre-training on comparable data and/or the close similarity between the objects of interest, i.e., craters and mining pits, is the main cause of the reported high performance. In this thesis research it was attempted to transfer-learn VGG16 (the backbone CNN of our Faster R-CNN), pre-trained on remotely-sensed data from the Aerial Image Dataset (AID; Xia et al., 2017), on our own LiDAR dataset. The AID dataset consists of 10,000 annotated aerial images of 600 by 600 pixels with pixel resolution varying between 0.5–8 m. Contrary to our expectations, the resulting model performed worse than others that were pre-trained on the general purpose dataset ImageNet (Russakovsky et al., 2015) and further attempts were discontinued. Other research has also shown that CNNs pre-trained on ImageNet outperformed CNNs pre-trained on different remotely-sensed datasets in the classification of remotely-sensed data (Pires de Lima & Marfurt, 2019). This might be related to the sheer size of general purpose datasets such as ImageNet, which ranges in the millions of images, compared to available remotely-sensed datasets that contain tens of thousands of images (Pires de Lima & Marfurt, 2019). Besides, the successful application of transfer-learning in different earth observation domains (see Ma et al., 2019) indicates the ability of CNNs to adjust to different types of data, advocating the use of general purpose datasets for pre-training (Ball et al., 2017; Nogueira et al., 2017).

7.2.2 Progress

Considering the above challenges and the prerequisites for the applicability of automated detection methods for archaeological prospection, formulated at the start of this research (see Chapter 1.4), two different workflows have been developed: WODAN and CarcassonNet (Fig. 7.4). Both approaches use transfer-learning in which VGG16 (Simonyan & Zisserman, 2015) and ResNet34 (He et al., 2016) respectively, pre-trained on the general purpose dataset ImageNet (Russakovsky et al., 2015), are fine-tuned on our own developed training datasets of LiDAR images with annotated archaeological objects. WODAN is a multi-class detector of both discrete objects and landscape patterns, i.e., barrows, charcoal kilns, and Celtic fields, while CarcassonNet has been specifically developed to deal with the problem of irregular landscape patterns, i.e., medieval hollow roads.

Both workflows solely make use of open-source data and software: the utilized LiDAR and other geospatial data is freely available from the online spatial data repository PDOK (Nationaal Georegister, 2021). The LiDAR data has been processed with *QGIS* (QGIS Development Team, 2017) and visualized with the *Relief Visualization Toolbox 1.3* (Kokalj & Hesse, 2017). Preprocessing, post-processing, and LBR are implemented in *QGIS* (QGIS Development Team, 2017) and *Python* (Van Rossum & Drake, 2009). Faster R-CNN is written in *Python* using the *Keras* library (Chollet, 2015) on top of Tensorflow (Abadi et al., n.d.), while ResNet34 is written in *Python* using the *Fast.ai* library (Howard & Gugger, 2020) on top of Facebook's PyTorch Artificial Intelligence development framework (Paszke et al., 2017).

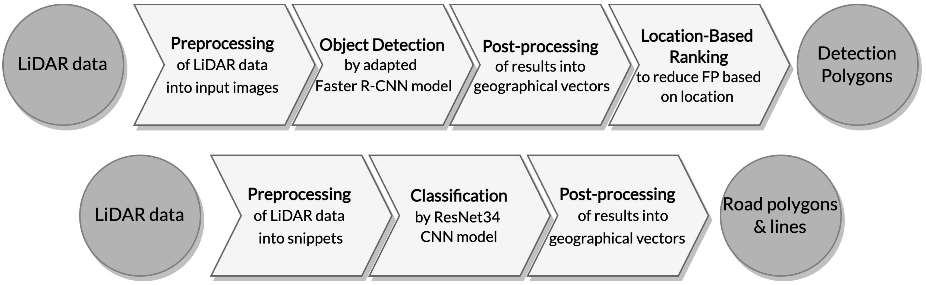
LiDAR data
Preprocessing of LiDAR data into input images
Object Detection by adapted Faster R-CNN model
Post-processing of results into geographical vectors
Location-Based Ranking to reduce FP based on location
Detection Polygons
LiDAR data
Preprocessing of LiDAR data into snippets
Classification by ResNet34 CNN model
Post-processing of results into geographical vectors
Road polygons & lines

Figure 7.4: Simplified representations of the WODAN2.5 and CarcassonNet workflows.

WODAN

The final version of the WODAN workflow, WODAN2.5 (Chapter 3.8), is a multi-class detector able to detect barrows, Celtic fields, and charcoal kilns, in both a small, non-random, and a large, random test dataset (see Chapter 3). The workflow has also been successfully applied 'in the wild', in a case study in the southern Netherlands (see Chapter 4). WODAN2.5 consists of four steps (Fig. 7.4): 1) A preprocessing step that 'cleans', visualizes, cuts, and normalizes the LiDAR data into input images;
2) an object detection step that uses an adapted version of the Faster R-CNN architecture (Ren et al., 2017); 3) a post-processing step that converts the output of the prior step, i.e., bounding boxes with pixel coordinates, a class label, and a confidence score, into geospatial vectors (either points or polygons) with real-world coordinates,

directly usable in GIS; and 4) an additional post-processing step called Location-Based Ranking (see Section 7.2.3) that incorporates domain knowledge into the workflow.

In the WODAN2.5 workflow different modifications were made to Faster R-CNN to cope with the above mentioned challenges and to adapt to the specific archaeological task. To address the problem of small objects, a relatively simple strategy was used that reduces the window in which the algorithm looks for objects. This is possible by lowering the size of the anchor boxes (see Chapter 3.3.3) generated by the Region Proposal Network (RPN; Ren et al., 2017) in Faster R-CNN, based on the approximate size of the objects of interest (Chen et al., 2017; Ren et al., 2018b; Tang et al., 2017b). This strategy was chosen over other techniques, such as magnifying the input images, combining the output of multiple layers of the neural network, i.e., multilevel feature fusion, or using multiple detectors with multiple scales (Guyot et al., 2018; Van Etten, 2018), as these result in a more complex model and a considerable increase in computational cost (Ren et al., 2018b). The problem of inconsistent orientation of objects can be dealt with by using data augmentation (Van Etten, 2018). In this research image flip (horizontally and vertically) and rotations were implemented (see Chapters 2 & 3). An additional benefit of data augmentation is that the dataset is effectively multiplied, reducing overfitting (Goodfellow et al., 2016). To reduce the foreground-background class imbalance the Focal Loss function (Lin et al., 2020) was applied in the RPN (see Chapter 3.8).

The issue related to the large size of the LiDAR images, as compared to images in general purpose datasets, has been addressed by cutting the images into subtiles of 600 by 600 pixels with an overlap of 30 pixels to all sides (see Chapter 3.3.2). The latter eliminates potential edge effects resulting from the visualization of the LiDAR data, and avoids the dissecting of archaeological objects on the edges of subtiles. By normalizing the subtiles, such that each image has pixel (or grayscale) values between 0 and 255, any detrimental effects of the dissecting of the data and potential problems due to relative height differences to the training of the CNN are negated (see Kazimi et al., 2019).

To successfully detect Celtic fields, a novel approach was taken to annotate these landscape patterns: instead of labeling the entire area as a single example, individual plots within the Celtic field were annotated as individual objects (see Chapter 2). This not only considerably increases the number of examples in the training dataset, but also proves to be a suitable solution to the problem of detecting Celtic fields. A downside of this approach is that instead of looking for a checkerboard pattern, which has few parallels in the landscape (see also Risbøl et al., 2013), WODAN looks for square or rectangular embankments, a shape much more abundant in the landscape. This results in more False Positives being caused by comparable objects.

CarcassonNet

The CarcassonNet workflow uses a combination of a Deep CNN and geospatial and image processing algorithms to detect and trace hollow roads (i.e., irregular landscape patterns) in LiDAR. CarcassonNet has been successfully applied on LiDAR data from the Veluwe (see Chapter 5), as well as on data from Germany and Slovenia (see Chapter 6). The workflow has a comparable structure as WODAN but consists of only three steps (see Fig. 7.4): 1) A pre-processing step that visualizes, cuts, and normalizes the LiDAR data into input images; 2) a classification step that uses the ResNet34 architecture (He et al., 2016) and Location-Based Ranking; and 3) a post-processing step that converts the output of the CNN into two types of geospatial vectors to efficiently study the roads themselves and their precise location in the landscape (polygons), and the course of the roads and the resulting route network (lines; see Chapter 5.2.2).

The most important contribution of CarcassonNet to the above challenges is a solution to detecting irregular landscape patterns. Comparable to the approach taken for the detection of Celtic fields (see above), CarcassonNet uses individual sections of roads as input, instead of whole roads. Therefore multiple pieces per single hollow road can be taken, making it much more cost-effective to create a sufficient training dataset. Furthermore, other challenges that are specifically related to object detection, such as the size and distribution of the objects, are resolved in CarcassonNet by separating the characterization and localization problem (see Chapter 1.2.2). By reducing the characterization sub-task to a binary classification, a CNN is only used for classification, a relatively simple task. This results in better performance at a lower cost/effort (Guo et al., 2016). Moving the localization sub-task to the post-processing step offers the opportunity to fully employ the capabilities of GIS, making it easier to process the output of the classification into results usable for archaeology. Apart from this, CarcassonNet addresses the above challenges by using overlap in the splitting of the LiDAR data, normalization of the resulting snippets, and by balancing the training dataset through data pruning (Angelova et al., 2005) and down and upsampling (He & Garcia, 2009).

7.2.3 Techniques to Add Domain Knowledge

To make automated detection methods usable in large-scale archaeological survey, additional information to differentiate between these morphologically identical objects needs to be incorporated into the classification process. Within archaeological object detection, different methods (see Table 7.1) have been developed that add this information, also called domain knowledge (Chapter 3.4, Davis et al., 2018; Meyer et al., 2019; Meyer-Heß, 2020; Zingman, 2016; Zingman et al., 2014). The basic assumption of these techniques is that the location of archaeological objects or objects of confusion in the landscape is not random but is, among others, the result of certain characteristics of the past and present environment (see also Verhagen & Whitley, 2020). Objects of confusion do not appear randomly throughout the landscape, but are generally related to specific natural or anthropogenic phenomena, such as drift-sand areas and modern roads.

Furthermore, although the location of archaeological objects is often related to their topographical or environmental context (Verhagen, 2007), e.g., the proximity of Viking age fortresses to watercourses (Stott et al., 2019), the current distribution of archaeological objects and their visibility in remotely-sensed data is generally the result of variable preservation or ground visibility conditions in different parts of the landscape (Bourgeois, 2013; Casarotto et al., 2018). The developed methods (Table 7.1) use this information about these so-called map formation processes (Fokkens, 1998) to enhance performance of automated detection methods.

Table 7.1: Approaches to add domain knowledge to reduce false positives.

Research	Method	Step	Results
Meyer et al. (2019)	Digital Landscape Model (DLM)	preprocessing	exclude areas that are exposed to anthropogenic relief-changing activities from detection
Zingman et al. (2014)	Morphological Texture Contrast (MTC) descriptor	preprocessing	exclude high contrast texture regions (urban areas, forests, rocky mountains) from detection
Davis et al. (2018)	land-use maps	post-processing	exclude detections in proximity to certain landscape elements
Verschoof-van der Vaart et al. (2020)	Location-Based Ranking (LBR)	post-processing	rank detections based on their location in relation to certain landscape elements

A general strategy of these methods is to either exclude specific parts of the landscape from classification or exclude detections made in these parts (Davis et al., 2018; Meyer et al., 2019; Zingman et al., 2014, 2016). For instance, Meyer et al. (2019) filter out areas that were (or still are) exposed to (sub)modern anthropogenic relief-changing activities, e.g., urban development, as it is unlikely that archaeological objects have been preserved in these parts of the landscape (Meyer-Heß, 2020). An alternative strategy is to add a specific 'terrain' class to the train and test datasets, in order to reduce confusion between objects of comparable morphology (Somrak et al., 2020; Trier et al., 2019). Location-Based Ranking (LBR), developed in this thesis (see Chapter 3.4), differs from these methods in that it does not *a priori* exclude any areas or detections from the classification process. Instead, detections are ranked according to their location in relation to specific landscape characteristics. Whether subsequently all detections, or a selection of detections based on their ranks, are used for further analysis is up to the operator and might depend on the goal of the survey (Chapter 3.6). Contrary, the approach taken by Stott et al. (2019) actually considers the topographical context, i.e., the proximity of possible archaeological objects to bodies of water, roads, and present-day place names or toponyms, in the classification process.

The methods in Table 7.1 have proven effective in adding domain knowledge to archaeological automated object detection.

For instance, the use of the Digital Landscape Model in Westphalia (Germany) reduced the amount of False Positives with circa 35% (Meyer-Heß, 2020), while the implementation of LBR on the Veluwe resulted in a reduction of False Positives with up to 70% (see Chapters 3, 4 & 5). Nevertheless, the development and successful implementation of the majority of these approaches is highly dependent on the availability of the required geospatial data (but see Zingman et al., 2014). While the Netherlands has a wide variety of high-quality geospatial data readily available (Nationaal Georegister, 2021), the level of availability, coverage, and quality varies widely between countries (Opitz & Herrmann, 2018). Of course, the latter is of influence on the usefulness of these methods, as the quality of the approach is limited by the data from which it was generated (Meyer-Heß, 2020).

While these approaches have a comparable goal, there is a clear methodological difference between LBR (see Chapter 3.4) and the strategies proposed in other work that exclude portions of the terrain or detections made therein from the results (Davis et al., 2018; Meyer et al., 2019; Meyer-Heß, 2020; Zingman, 2016). These 'exclusion strategies' have, in addition to the reduction of False Positives, the advantage that reducing the investigated area also decreases computational costs and processing time (Meyer-Heß, 2020). However, their implementation has a potential risk: it can result in a self-fulfilling feedback system (Wheatley, 2004). As these strategies are based on the known distribution of archaeological sites, they effectively reproduce and reinforce the inherent biases within the existing archaeological record (Nuninger et al., 2020b). Consequently, using exclusion strategies results in the systematic searching for undiscovered sites in places where we expect to find them. And if we only look in those places, we will only find them there (Nickerson, 1972). In contrast, LBR offers the opportunity to analyze the detections made in less-favorable areas. This can lead to new insights on the preservation conditions of specific areas, which in turn can result in changes in subsequent versions of the method. For instance, initially modern roads were considered as detrimental to the preservation of all archaeological objects on the Veluwe and were given a low rank on the LBR maps (see Chapter 3.4.1). However, during the preliminary analysis it was noted that while many Celtic fields were intersected by roads, this has had a limited negative impact on the preservation of the overall objects. On the other hand, discrete objects, e.g., barrows, were severely damaged or even completely destroyed by roads. Therefore, the LBR map was adjusted and roads were given a high rank in the case of Celtic fields, while for discrete objects the lower rank was maintained. A further complication of exclusion strategies lies in multi-class detection, in which the classes vary in date and/or have a different preservation in particular areas. For instance, the majority of large-scale drift-sands on the Veluwe originate from the Middle Ages (Koster, 2009). Obviously, these areas are detrimental for the preservation and visibility of prehistoric traces, such as barrows. On the other hand, charcoal kilns, dating from the late Middle Ages or later, are not or hardly affected by these drift-sand areas. The binomial nature of exclusion strategies does not take such discrepancies or intricacies into account, unless for every class the exclusion strategy is adjusted and the dataset or detection model is reimplemented, which decreases efficiency and negates the advantage of reduced computational costs.

Improving Domain Knowledge-including Techniques

Current strategies to include domain knowledge to automated detection are based on the relation between archaeological object classes and specific anthropogenic and natural landscape elements, for example barrows and drift-sand areas (see Chapter 3.4). These methods could be further improved and extended by incorporating information on the spatial relationship or dependence among objects within a specific class or between different classes, i.e. spatio-contextual information (Li et al., 2014). For instance, in this thesis individual plots within a Celtic field are detected as opposed to the entire Celtic field. Therefore, clusters of multiple detections, as opposed to single, isolated detections, are much more likely to actually indicate a Celtic field. This concept could be used to further reduce False Positives by disregarding isolated detections and only look at clusters. However, this requires a high level of completeness—how accurately the results reflect the extent of the archaeological objects (see Chapter 4.4.1). If the completeness of a model is sub-optimal, such as in the case study in the *Midden Limburg* area (see Chapter 4), disregarding isolated detections could actually be detrimental to the overall performance.

Alternatively, the confidence score of detections could be penalized or increased based on the proximity of other detections of the same class. For instance, bomb craters from the Second World War generally appear in (linear) clusters due to the purpose, i.e., destroying a particular enemy target, and the practicality, i.e., the use of so-called bomb runs, of aerial bombardment (see for instance Passmore et al., 2014; Waga & Fajer, 2021). Therefore, in an approach that detects these craters, detections could be given an increase to their confidence scores based on the spatial proximity to other detections, while isolated detections could be penalized (see also Brenner et al., 2018). Furthermore, in multi-class detection the presence of bomb craters could be used as an indication that other traces of conflict might be present, i.e, the target of the aerial bombardment, and detections of other classes could be given an increase in confidence when bomb craters detections appear in close proximity. Although, if the bombardment was successful, the target was most likely (partially) destroyed and might not be preserved.

Further potential for improvement lies in the addition of expert cognitive processes and reasoning to the classification process. As said, human interpreters can often make a distinction between objects of confusion and archaeological objects. These interpretations are often based on a complex mix of prior experience, knowledge, and a degree of flexible and creative reasoning that are almost impossible to untangle (Halliday, 2013). But deconstructing these often subconscious processes could help us to better understand the reasoning underlying an expert's definition and identification of archaeological objects or objects of confusion (Bennett et al., 2014; Nuninger et al., 2020b). Unfortunately, explicit verbalizing of these thought processes and reasoning, i.e., elicitation, has proven difficult (White, 2019). This is further complicated by semantic inconsistency (Davis, 2020). Moreover, among interpreters there is a variability in detection accuracy (Risbøl et al., 2013; Sadr, 2016), leading to different interpretations of the same data which are often highly personal (Quintus et al., 2017). Consequently, these thought processes and reasoning are still poorly understood, in-explicit, and not reproducible (Cowley et al., 2020; Davis, 2020).

In that sense, humans are as much 'black boxes' as is sometimes claimed of CNNs (see Castelvecchi, 2016). Yet, a better understanding and quantification of expert reasoning would greatly improve domain knowledge including techniques (White, 2019).

7.2.4 Solutions for Automated Detection in Archaeology

The experiments conducted with WODAN and CarcassonNet have shown the potential of (Region-based) CNNs for archaeological object detection. This potential is further demonstrated by an abundance of other research projects that report favorable performance of CNNs on the classification, object detection, and segmentation of archaeological objects in remotely-sensed data (see Chapter 1, Fig. 1.2 for an overview). However, the implementation of these (traditional) object detection approaches is limited by the above mentioned challenges. In this research these challenges have been addressed by using an established object detection model (Faster R-CNN), to which different modifications and extensions (e.g., anchor box reduction, Focal Loss, and Location-Based Ranking) were added. The different experiments have shown that both WODAN and CarcassonNet work adequately, although their performance still has room for improvement. For instance, WODAN has not reached general human performance on the same task (see Chapter 3).

Concerning the different types of detection used in this research it can be concluded that for multi-class detection problems, especially when different object classes appear in close proximity or even overlap, object detection is the preferred method (see also Fiorucci et al., n.d.). In the case of a detection problem involving a single object class it is more beneficial to split the classification and localization sub-tasks, as was done in CarcassonNet. This is possible by combining a 'normal' CNN for classification and employing GIS in the post processing for localization (see Chapter 5.2.2). Especially for irregular landscape patterns (e.g., roads) this method has proven useful as it removes the problem of producing fitting bounding boxes. The question remains what level of performance is needed for these methods to be confidently employed in archaeological practice. To determine this, the actual evaluation of the performance of archaeological detection methods needs to be reviewed to facilitate comparisons between methods.

7.3 Evaluation of the Performance of Detection Methods in Archaeology

Arguably, one of the most important parts of many archaeological detection research projects is the assessment of the performance of the developed method, and a comparison of the performance with other approaches (see for instance Kazimi et al., 2020a; Trier et al., 2019). However, as shown in Chapters 3 and 5 the performance of a method is in large part dependent on the datasets, metrics, and evaluation methods used. These same factors also make a fair comparison between different methods difficult (see Chapter 2.6). Therefore, in the following these different aspects will be discussed in more detail.

7.3.1 Datasets

Generally, in archaeological detection research considerable attention is paid to the preprocessing of remotely-sensed data and the creation of the training dataset (see Trier et al., 2019). For instance, multiple research projects have focused on evaluating different LIDAR visualization techniques to be used for input images of CNNs (Gallwey et al., 2019; Guyot et al., 2021a; Kazimi et al., 2020b; Somrak et al., 2020). This preprocessing is certainly of importance, as the training data is a key-component of Deep Learning approaches. Deep CNNs are data-hungry beasts—as opposed to Machine Learning applications that often perform well even if the dataset is small—of which the performance is in large parts linked to the amount and quality of the data presented to it. The more data a CNN is given, the better it is likely to perform (Kumar & Manash, 2019). And obviously, these approaches can only be as good as the data from which they are created (Chapman, 2011). The relation between the size of the training dataset and performance could be observed during this research: with only 177 charcoal kiln examples in the initial dataset, WODAN was unable to detect this class (Chapter 2). In subsequent research the dataset was enlarged with circa 400 additional charcoal kiln examples, which proved sufficient for the model to detect charcoal kilns, albeit with a relative low performance (Chapter 3). Therefore, a logical way to further improve the performance of WODAN—and other Deep Learning approaches—is by increasing the number of examples in the training dataset. Yet, expanding and combining archaeological training datasets brings about several issues, such as veracity (i.e., differences in quality of information between datasets), incompleteness, and the issue of using unverified data (McCoy, 2017).

With the aggregation of data from different and heterogeneous providence, a certain degree of messiness should be expected and is probably inevitable (Gattiglia, 2015). This messiness results from combining, extracting, and transforming data, inconsistencies in formatting, and unavoidable (human) errors. In the case of geospatial or remotely-sensed data this generally emerges as variations in positional accuracy or measurement quality, although semantic inconsistency can also become problematic (Davis, 2020; Gattiglia, 2015). Furthermore, archaeological datasets are inherently incomplete and almost always contain a certain degree of unknown archaeology (Opitz & Herrmann, 2018). For example, during the creation of the datasets used in this research, over 700 new potential barrows were discovered, in an area that was regarded as extensively surveyed (Lambers et al., 2019). Moreover, in the development of CarcassonNet (Chapter 5.3.3), a feedback loop between the archaeological interpreter and the classification algorithm showed that in the initial manual labeling of the dataset circa 5.5% of the hollow roads were missed. It is also regularly noted in various research that False Positives from automated detection methods could actually be unlabeled or unconfirmed archaeological objects in the test dataset, so called 'New Positives' (Bonhage et al., 2021; Meyer et al., 2019; Trier et al., 2019). This complexity and inconsistency in datasets constrains the development and performance of automated detection methods (Casana, 2020; Lambers et al., 2019; Sadr, 2016). Especially for approaches that use fully automated preprocessing to create training data this can cause problems (see for example Olivier & Verschoof-van der Vaart, 2021).

Finally, expanding datasets with unverified archaeological objects remains a point of discussion. Archaeologists generally make a clear (quantitative) distinction between verified and unverified objects or sites, and put much reliance into field-validated sites, while unverified sites are generally distrusted and given low confidence (Banaszek et al., 2018; Cowley, 2012). However, only using fully field-validated data is both financially and practically unfeasible, although efforts that utilize Citizen Science are a potential solution (see Lambers et al., 2019). In other domains the use of unverified sources of information is justified by the logic that the sheer volume of data will overcome the inclusion of some poor quality data (McCoy, 2017). Furthermore, enlarging the training dataset with unverified archaeological objects, which often are not consolidated or reconstructed—and therefore better resemble undiscovered objects—might actually benefit the generalization capabilities of the approach (see Chapter 3.4). Besides, it could also be argued that for certain types of archaeology (e.g., Celtic fields or hollow roads) remotely-sensed data might be the optimal medium to evaluate the presence of these objects, as validation in the field is problematic due to the lack of clear traces and material culture (Lambers et al., 2019). Therefore, the benefits of adding unverified data might outweigh the drawbacks.

While the training dataset is in general carefully created, the importance of the composition of the test dataset is often overlooked. As shown in Chapter 3, the performance of the same model can vary significantly between test datasets in which the state of preservation of the archaeological objects varies and that have a different density of archaeological objects, i.e., a different ratio of positive and negative examples. For example, Brenner et al. (2018) demonstrate a Precision (see Fig. 7.5) of circa 90% in the detection of bomb craters in historic aerial photographs on a test dataset with an 1:1 ratio of positive and negative examples. However, they also note that in a realistic scenario the ratio of positive and negative examples is approximately 1:250, and the Precision would drop to 4%. Therefore, a distinction can be made between experimental and 'in the wild' testing. In the former the test dataset is often small, non-random, and selective, i.e., it is an excerpt of the available data often containing well-visible or well-preserved objects and/or a disproportionate number of positive examples. Contrary, testing in the wild concerns a large dataset of all available data from a certain area (and therefore a realistic distribution of positive and negative examples). These datasets generally also include obscured objects or objects in a bad state of preservation (see Chapter 3). Both types of test dataset have merit in different situations. Testing on an experimental test dataset gives a good indication whether the applied method is suitable, on a technical level, for the specific task. These datasets are usable to test newly developed approaches (e.g., proof of concepts) or to assess the improvement of additional measures, e.g., new data augmentation or loss functions, in an existing method. On the other hand, a random test dataset better represents the real-world situation of the prospection of scarce archaeological objects over different types of complex terrain, and therefore gives a better indication of the practical value of the automated detection model for archaeological practice. Therefore, a well thought-out composition of the test dataset or datasets is essential for representative results.

7.3.2 Metrics

Within archaeological automated detection a variety of performance metrics are used to evaluate methods (see for instance Bundzel et al., 2020). The majority of studies use metrics computed from a confusion matrix (Gong, 2021), i.e., Accuracy or a combination of Precision, Recall, and F-score, generally F1 (see Fig. 7.5). Accuracy gives the ratio of correctly predicted instances to the total instances in the dataset. The F1-score is the harmonic mean of the Precision and Recall and gives a measure of a model's performance (Sammut & Webb, 2010). This metric can also be used in multi-class detection, by using the macro/micro averaged F1-score (Chicco & Jurman, 2020; see Chapter 2). It should be noted that during the actual training of Deep CNNs rather than these metrics, the loss function—a function that calculates the penalties of incorrect classifications into a single number (Goodfellow et al., 2016)—is optimized. A low loss function is generally regarded as an indication for a well-trained approach and therefore high performance (Guo et al., 2016).

		Actual values	
		Positive (1)	Negative (0)
Predicted values	Positive (1)	TP	FP
	Negative (0)	FN	TN

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

$$F1 = 2 \times \frac{Recall \times Precision}{Recall + Precision} \qquad ACC = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TN + FP) \times (TN + FN)}} \qquad nMCC = \frac{MCC + 1}{2}$$

Figure 7.5: Overview of different metrics generally used in archaeological automated detection research.

Both Accuracy and F1-score are among the most popular statistical measures in Machine Learning (Chicco & Jurman, 2020). However, the use of these metrics can in certain circumstances result in overoptimistic or sub-optimal results, especially when imbalanced datasets are used (He & Garcia, 2009; Johnson & Khoshgoftaar, 2019; López et al., 2013; Luque et al., 2019; Santos et al., 2018). Class imbalance—when the number of examples in one class is much larger than the number of examples in the other class(es)—is naturally inherent in many real-world situations, such as in most archaeological datasets used for automated detection. This is especially the case in 'realistic' test datasets, where the distribution of positive and negative examples is very skewed (see Section 7.3.1). Usually, the minority class represents the concepts of interest, e.g., archaeological objects, whereas the other class (the majority class) represents the counterpart of that concept, e.g., empty terrain.

When a dataset is imbalanced, Accuracy becomes an unreliable metric (see for instance Chapter 5, Table 5.5), since it does not distinguish between the number of correctly classified examples of different classes, and therefore provides an overoptimistic estimation of the classifier's ability on the majority class (Chicco & Jurman, 2020; López et al., 2013). Simply said, an approach that classifies every example as the majority class, will reach a high Accuracy, due to the overabundance of these examples compared to the minority class(es). For example, if only 1% of the instances in a dataset belong to the minority class, a classifier that always outputs the majority class label for all instances will achieve an Accuracy score of 99%! This high score misleadingly indicates good performance, while the approach will be useless in practice (Johnson & Khoshgoftaar, 2019). Contrary, Precision and Recall, and consequently the F-score (see Fig. 7.5), only consider positive examples and predictions, generally the minority class(es), and do not show how the approach handles correct negative examples (True Negatives, TN; see Fig. 7.5), which is usually the majority class (Chicco & Jurman, 2020; Gong, 2021). Therefore, these metrics are not affected by class imbalance, but on the other hand do not show how well an approach can identify negative examples. Furthermore, the F1-score is very sensitive towards disparate values, i.e., outlying scores in Precision or Recall, which can lead to a sub-optimal approach (López et al., 2013). Related to this is the fact that the F1-score assigns equal importance to both Precision and Recall. However, this balance might not be desirable as the relative importance of Precision and Recall is an aspect of the task of the method. For instance, as argued in Chapter 3.6, within archaeology the preferred balance between Precision and Recall depends partly on the (envisioned) users and the task of the method. A high Recall might be preferred when localizing as many of the archaeological objects as possible is paramount for appropriate conservation and heritage management. Contrary, when limited resources are at hand for (field) validation, a high Precision is preferred. Generally, priority is given to Recall in archaeological research (see for example Soroush et al., 2020). Although, completely neglecting Precision runs the risk of moving the professional bottleneck (Smith, 2014) from the desk to the field, i.e., the main problem becomes (field) validating, instead of detecting, an overwhelming amount of potential archaeological objects. Therefore, other versions of the F-score, where the '1' is replaced for either '2' or '0.5' in the equation might be more fitting for archaeology.

The former (F2-score) weighs Recall higher than Precision, while the latter (F05-score) weighs Recall lower than Precision (Sammut & Webb, 2010). Of course the use of alternative versions of the F-score should be clearly stated to avoid confusion.

Balancing Solutions

As the main problem of overoptimistic or sub-optimal metrics lies in the imbalance between the majority and minority classes, different solutions on either a data, algorithm, or metric level have been developed to resolve this (Oksuz et al., 2019). Solutions on a data level generally involve sampling methods to modify an imbalanced dataset into a balanced distribution of positive and negative examples, i.e., random over and undersampling (He & Garcia, 2009). The former duplicates random examples from the minority class, while the later discards random examples from the majority class. Over and under-sampling have been successfully implemented to balance the training dataset of CarcassonNet (see Chapter 5.2.2). However, these techniques should be used with caution because random over-sampling might cause overfitting, while under-sampling reduces the amount of information in the dataset from which the approach has to learn. Consequently, a variety of additional 'intelligent' methods have been developed that do not randomly over or under-sample the data, but attempt to either reduce overfitting or preserve information for learning (Johnson & Khoshgoftaar, 2019). Solutions on an algorithm level generally concern new loss functions, e.g., Focal Loss (see Section 7.2.2 and Chapter 3.8), different training schemes, and threshold moving (Johnson & Khoshgoftaar, 2019; Zou et al., 2016).

An effective metric to evaluate approaches that use an imbalanced dataset is Matthews Correlation Coefficient (MCC; Chicco & Jurman, 2020; see Fig. 7.5). Contrary to Accuracy and F1-score, MCC is not bound between 0 and 1 but between -1 and 1. In order to obtain a MCC between 0 and 1, normalized MCC (nMCC; Fig. 7.5) can be used (Chicco & Jurman, 2020). Comparable to F1-score, MCC can also be calculated for multiple classes (Grandini et al., 2020). MCC is regarded as a balanced metric, which takes all aspects of the confusion matrix (i.e., TP, FP, TN, and FN) into account. Thus to get a high MCC score, a classifier has to make correct predictions both on the negative and positive cases, independently of their ratio in the overall dataset (Chicco & Jurman, 2020). For this reason, it is recommended to use either MCC (see Chapter 5.3) or, preferably, a combination of different metrics including MCC (see for instance Bundzel et al., 2020) to evaluate performance of archaeological detection methods. Reporting results using a combination of diverse metrics also facilitates and simplifies the comparison of different methods.

7.3.3 Evaluation Methods

Apart from the usefulness of different metrics for archaeological detection methods, it might be considered whether the evaluation methods used are actually suitable for our specific task (see also Fiorucci et al., n.d.). In other words: Do we need to evaluate how we evaluate our methods? For instance, Gallwey et al. (2019) show that even if individual detections are not always correct, the overall patterns in the landscape are correctly reproduced by their method.

A comparable issue is raised in Chapter 4, where the matter of completeness—how accurately the results reflect the extent of the archaeological objects—is discussed. While WODAN is able to detect the majority of demarcated areas of Celtic fields within the research area, it generally does not detect the full extent. This results in low(er) performance when calculating metrics based on coverage, even though the method did correctly point to the location of archaeological objects and shows the overall pattern of Celtic fields in the area. Therefore, the proximity of a detection to the actual location of an archaeological object might be more important than the exact intersection or coverage for archaeological object detection. Bundzel et al. (2020) have quantified this with the MOR10R metric, which calculates the number of misclassified pixels further than 10 m from True Positives over all pixels. The concept is to minimize the proportion of misclassified pixels outside an area that will be inspected by an expert, which is near the True Positives. Basically, errors close to True Positives are less detrimental than those farther away (Bundzel et al., 2020). These notions are in line with Cowley's idea of a conceptual shift from a widely held fixation on individual detections being correct to the overall patterns being descriptive (Cowley, 2012; Sadr, 2016). While the correct detection of individual objects is most beneficial for settlement archaeology, correctly identifying the overall pattern is especially relevant for environmental and landscape archaeology (see Chapter 1.1). Considering the incorporation of these methods in archaeological practice, quantifying the 'overall pattern descriptiveness' of a method would be valuable information. In the research of Sadr (2016) patterns, resulting from the classification of settlements by different interpreters, are compared by visually inspecting grayscale heatmaps. This methodology could be translated into an evaluation metric, e.g., by using statistical tests developed in Computer Vision to compare grayscale images (Wilson et al., 1997) or in Ecological research to compare the spread of populations (Syrjala, 1996), to compare the known distribution of archaeological objects in an area to the distribution resulting from using automated detection methods.

7.3.4 Transferability

Next to the performance on a random test dataset, an important factor determining the applicability of an automated detection method is their transferability (Cowley et al., 2020; Kermit et al., 2018), i.e., the usability of the method on an unrelated area with either different topography, land-use, and/or LiDAR data of different properties. As can be imagined, these factors can vary considerably on a regional or national 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 for archaeological practice. In Chapters 4 and 6, the transferability of WODAN and CarcassonNet has been investigated. WODAN was trained on data from the central part of the Netherlands (the Veluwe) and subsequently applied to data from the southern Netherlands (Midden Limburg). While the parameters of the LiDAR data (e.g., average point-density, resolution, etc.) were the same for both regions, these areas varied in archaeological, geo(morpho)logical, and land-use conditions (Chapter 4.2.1). Contrary, in the transferability study of CarcassonNet, the topography, land-use, as well as the properties of the LiDAR data varied (Chapter 6.2.1).

Both studies showed that the approaches are able to generalize to different situations, although in general the performance on the new datasets decreases. The influence of different terrain and land-use seems minor. This is probably due to the fact that the majority of the possible terrain and land-use encountered is present within the training datasets in some degree. Additionally, issues with particular terrain and land-use can be overcome by adding (additional) domain knowledge (see Section 7.2.3). However, differences in the properties of the LiDAR data, especially the average point-density, seem to be detrimental to the ability of methods to detect archaeological objects (Chapter 6.4). This has also been observed in other research on the automated detection (Dolejš et al., 2020; Trier & Pilø, 2012) and the manual analysis (Risbøl et al., 2013) of archaeological objects in LiDAR data.

Related to this is the question whether transferability should be expanded with the ability of a particular method or workflow to be (re)applied to different types of remotely-sensed data and/or archaeological objects. For instance, is it beneficial to fine-tune WODAN, which already has been trained to detect prehistoric traces, to detect traces of conflict? Or does this involve the complete retraining and restructuring of the workflow? It would be beneficial to focus future research on the development of general workflows and methods that can easily be modified and fine-tuned on particular data and archaeological objects.

7.4 Incorporating Automated Detection in Archaeological Practice

The research in this thesis has shown the potential of workflows like WODAN and CarcassonNet to detect different types of archaeological objects in LiDAR data. In the preceding sections the progress made towards resolving the challenges of using (R-)CNNs for archaeological automated detection has been discussed. Furthermore, the datasets, performance evaluation, and metrics used for these methods have been reviewed. This all leads to the discussion on how automated detection methods can be incorporated in archaeological practice, which is the aspiration of many research endeavors (see Kermit et al., 2018; Trier et al., 2019), including this one. To date only a few automated detection methods have actually been applied in archaeological practice or heritage management (see Kermit et al., 2018). According to Opitz & Herrmann (2018) this lack of incorporation is prompted on a technological level by the fact that these approaches are still in an early stage of development with unsatisfactory results. On a practical level the minimal requirements of automated detection methods for specific activities within archaeological practice (e.g., decision making, planning, and research) remain undefined (Opitz & Herrmann, 2018). In my opinion these two factors are directly related. To move from a development stage to an application stage, initially a baseline of acceptable performance needs to be established (see Cowley et al., 2020; Stallkamp et al., 2012).

7.4.1 Setting the Bar for Archaeological Automated Detection

Current automated detection research focuses mainly on the development and assessment of methods within a vacuum, as the field lacks consensus on the minimal acceptable performance for practical use—while benchmarks to assess this performance are also absent (see Section 7.6.4). This problem is not limited to archaeology, but is prevalent in many domains in which Deep Learning is used (see Zhang et al., 2021). Some archaeological research describes the minimal level of performance in general terms, such as that methods need to produce 'reasonable' low numbers of False Positives and False Negatives to be applicable (see for instance Kermit et al., 2018). In other cases, the minimal level of performance is equated with the capabilities of human experts to analyze remotely-sensed data (Casana, 2020). This expert performance is generally regarded as exceptionally good, to the point that the results of their manual analysis are essentially taken as truth (Raczkowski, 2020), even though classifications by different interpreters are highly subjective and can vary widely (Quintus et al., 2017; Risbøl et al., 2013; Sadr, 2016). This has in part fueled the distrust towards automated detection methods, as the performance of these often fall behind manual mapping results (Traviglia et al., 2016). In my opinion, expecting expert performance from (current) archaeological detection methods is a pipe dream, as these are not only limited by their lack of expert knowledge, but also suffer from additional constrains on a methodological and data level (see Sections 7.2.1, 7.2.3, and 7.3.1). Although occasional reports are published about Deep Learning approaches outperforming humans (for instance Silver et al., 2016), these incidents generally concern (board)games, chemical or medical tasks, or relatively simple image classification tasks (Perrault et al., 2019). However, to date no CNN-based object detection approach, as used in this research, has reached superhuman performance (Zhang et al., 2021). At best, most Deep Learning approaches can approximate the performance of the humans that created their training datasets, due to the heavy reliance of these algorithms on humanly annotated data to effectively learn. These large datasets, which are an important part of the pre-training of CNNs, are generally not made by experts but through crowd-sourcing (Russakovsky et al., 2015). While these are certainly of high quality, they are neither exceptional nor error free (Northcutt et al., 2021). Besides, the comparison between the performance of human experts and automated detection methods often misses the point that the purpose of the latter is to rapidly detect potential archaeological objects in large areas as opposed to a detailed analysis and interpretation of all potential archaeology (Cowley, 2012). Consequently, expert's levels of performance as a minimal baseline for automated detection methods in archaeology seem unsuitable.

General Human Performance

Rather, if a minimal baseline should be set for automated detection methods, it could be argued that their performance should be more closely related to—and therefore can be better compared with—the capabilities of general humans. Because the latter share the same cognitive capabilities as experts to detect objects in images, but 'lack' the experience and expertise of expert interpreters (see also Cummings, 2014; Stallkamp et al., 2012).

Therefore, the performance of general humans better shows the achievable performance of automated detection methods, without expert knowledge—which so far has seem impossible to add to the classification process (see Section 7.2.3). Interestingly, a comparison between general humans, experts, and automated detection might actually inform us on which part of the variability in performance is related to the limitations in 'vision' and which part equals expert knowledge (see Section 7.2.1 and 7.2.3).

To asses this general human performance, methods and techniques from Citizen Science can be used (Fritz et al., 2017; Kosmala et al., 2016; Salk et al., 2016). The applicability of this type of research has been shown by recent projects in which archaeological objects were classified in remotely-sensed data by citizen researchers, generally without a background in either remote sensing or archaeology (Forest et al., 2020; Lambers et al., 2019; Lin et al., 2014a; Stewart et al., 2020). For instance, the results of the Heritage Quest project give a good indication of the performance of general humans in detecting archaeological objects in LiDAR images (Lambers et al., 2019). The results show that humans perform reasonably well on this task, although the performance varies between different types of archaeological objects (see Chapter 3). Interestingly, mistakes made by citizen researchers can often be related to objects of confusion (see Section 7.2.1), comparable to misclassifications made by automated detection methods. This underlines the issue of the lack of expert knowledge in both strategies (see Section 7.2.3). Aside, a high number of False Positives, often the main argument against automated detection methods (Casana, 2020; Traviglia et al., 2016), can also be observed in the results of the citizen researchers. Of course, the performance of the citizen researchers in the Heritage Quest project (Chapter 3.5.2) are specific to this particular research area and task, and might not be directly translatable to every other research project involving archaeological detection. The performance will differ when other types of archaeological objects are mapped, in different areas, and/or in other types of remotely-sensed data. Nonetheless, this general human performance serves as a better baseline for determining the achievable capability of automated detection methods than the performance of experts. A next step would be to develop benchmark datasets for archaeological automated detection, which also record the performance of general humans on that particular dataset (see Section 7.6.4). The question remains whether this human level of performance is actually necessary in all activities within archaeological practice, i.e., the specific role of the method within the broader archaeological (research) framework is of importance as well.

7.4.2 The Role of Automated Detection in Archaeological Practice

As argued in Chapters 3 and 4 the required performance of an automated detection method for incorporation into archaeological practice is dependent on the envisioned users, the intended task, and the embedding of the tool (and its results) within the wider archaeological research framework (see also Banaszek et al., 2018; Cowley et al., 2020; Lambers et al., 2019; Opitz & Herrmann, 2018).

Different users and tasks require an emphasis on either Recall, (i.e., maximizing the number of True Positives) or Precision (i.e., minimizing the number of False Positives; see Section 7.3.2). For example, the task of cultural heritage managers is often to evaluate and attach priorities to certain areas. This requires adequate information about the presence of archaeological objects within that area (i.e., a high Recall) to ensure appropriate conservation. On the other hand, a researcher studying the distribution of a certain archaeological phenomenon is more concerned with the results showing the overall pattern of objects, and therefore a high Precision might be preferred (see Chapter 3.6).

More importantly, when automated detection methods are used as a single source of information, without subsequent verification of the results, middling levels of performance are not adequate. However, using automated detection in this way seems neither scientifically sound, nor desirable (see also Bennett et al., 2014). Contrary, when automated detection methods are used in collaboration with other techniques, or are subsequently verified, moderate performance can be sufficient and practical (Opitz & Herrmann, 2018). Related to this is the amount of reliance that is put into the results of automated detection methods. As argued in Chapter 4, when the results of an automated detection method are taken as highlighting areas of interest, which contain potential archaeological objects that require (field) verification, rather than direct indicators of the presence of an archaeological object, the level of competence of the method does not have to be extremely high as the loss of a wrong detection is low (see also Opitz & Cowley, 2013). As noted by Kermit et al. (2018), the results of an imperfect automated detection method are better for aiding manual analysis or for selecting areas for field verification than no detections at all. Rather than thinking of automated detection methods as individual agents within archaeological prospection, the role reserved for automation on a complex task such as analyzing remotely-sensed data is as a teammate aiding humans in organizing, filtering, and synthesizing data (see Bennett et al., 2014; Cummings, 2014).

Human—Computer Collaboration

The concept of automated detection in a supplementary role, next to manual analysis, offers many opportunities for improving the investigation of remotely-sensed data (Bennett et al., 2014; Cowley, 2012; Trier & Pilø, 2012). This was shown in Chapter 4, where archaeological objects, missed during the manual analysis, were found by the automated detection method. The re-examination of the research area, guided by the results of the automated detection, resulted in even more potential archaeological objects, overlooked during both the initial manual analysis and the automated detection. It was shown that the interaction between human and computer (Boy, 2011), in which manual analysis and automated detection is combined in a so-called Human—Computer strategy, resulted in a more complete overview of the archaeology in the area and a gain in both quantitative and qualitative archaeological knowledge (see Cummings, 2014; Huggett, 2020a). The efficiency of these Human—Computer strategies lies in the fast run-time of automated detection methods, which offers opportunities to run multiple algorithms, which detect different types of archaeology, or run multiple versions of the same method simultaneously to improve performance (see Chapter 3).

An added benefit of Human—Computer strategies is the insight it can offer in the biases of both manual and automated analysis (Bennett et al., 2014; Halliday, 2013; Trier et al., 2019). It also resolves one of the caveats of current automated detection methods: these tools can only detect objects similar to the pre-defined target class(es) while other objects are ignored (Lambers et al., 2019).

Obviously, the level of human involvement—and consequently the performance of the automated method—in these strategies can vary based on the purpose of the research. To get a general baseline on the presence of certain archaeological objects, human involvement can be minimal, while an in-depth analysis of an area needs more involvement. Human involvement might also decrease over time, when the performance of the automated method increases, due to technical improvements or additional training data. Of course, a certain degree of involvement from a human interpreter remains necessary and desirable, as the purpose of automated methods is the detection of potential archaeological objects, not in the interpretation of these objects. The latter is and stays the prerogative of archaeological experts (see Traviglia et al., 2016).

7.4.3 Staunching the Data Deluge

From the preceding, it becomes obvious that I envisions the incorporation of automated detection methods in archaeological practice in a strictly supplementary role, preceding or in conjunction with manual analysis. These Human—Computer strategies, with various levels of human involvement depending on the task, resolve many of the current issues of archaeological detection and require only moderate levels of performance from developed methods. In my view, the incorporation of these approaches in archaeological prospection—to highlight areas of archaeological interest that require (field) verification, add detail to existing archaeological predictive maps, or to create the basis for fieldwork projects—lies in the very near future.

However, there are still some who deny the problems relating to the ever-increasing amount of remotely-sensed data, and/or argue against the use of automated detection in archaeology (Palmer, 2021; Parcak, 2009). According to them, the analysis of all data is not necessary, as years of experience have enabled experts to ignore data that is unlikely to serve our purpose, thereby effectively staunching the data deluge. Not only does such a strategy run the risk of turning in a self-fulfilling feedback system (Nuninger et al., 2020b; Wheatley, 2004), the fact remains that the analysis of remotely-sensed datasets—even when chunks of data are discarded or ignored—is a labor and time intensive effort, which often exceeds the limited resources available to archaeology (Lambers et al., 2019). Unsurprisingly, the analysis of remotely-sensed data on a regional, let alone national scale is hardly ever undertaken (Cowley et al., 2020; Hesse, 2013). The implementation of automated detection methods can at least alleviate part of the labor investment and provide a starting point for further (manual) analysis (Banaszek et al., 2018). A more pressing matter, necessitating automated detection, is the ever-increasing threat to archaeology around the globe.

While in most European countries this might comprise a slow rate of degradation due to modern land-use, e.g., agriculture, and urban development (Bonhage et al., 2021), in other parts of the world archaeology is under threat of extensive looting and systematic and deliberate destruction (El-Hajj, 2021). Even in an 'uneventful' region as the Dutch Veluwe, where little land development is taking place, archaeology is inadvertently damaged or destroyed (see for instance Fontijn et al., 2011). This is generally not deliberate but due to an unawareness of the presence of archaeological objects. The ability to rapidly detect objects in large remotely-sensed datasets might prevent this irretrievable loss of archaeological sites and information. Finally, these automated detection methods are not meant to replace expert interpreters or 'automate archaeology' (Traviglia et al., 2016). Rather, these methods can be valuable additions to the existing archaeologists' toolbox to rapidly map archaeological objects over extensive areas (see Soroush et al., 2020), to provide knowledge about human activity in the landscape, to improve the efficiency of remotely-sensed data analysis, and most importantly to reduce the expert's time invested in actually mapping objects, so that their time can be reallocated to analysis, validation, and interpretation.

7.5 Conclusions

To conclude, in this thesis the use of Deep (Region-based) CNNs for the detection of (multiple classes of) archaeological objects in remotely-sensed data, and the incorporation of these methods in archaeological practice was investigated. This research has shown the possibilities of Deep CNNs for this task, although the implementation of these architectures is limited by several factors, such as the nature and appearance of archaeological objects in remotely-sensed data, their similarity to other anthropogenic and natural landscape elements, and the characteristics of the remotely-sensed data itself. Two workflows have been developed in this thesis, named WODAN and CarcassonNet, which combine CNNs and GIS to detect barrows, Celtic fields, charcoal kilns, and hollow roads in LiDAR data. Location-Based Ranking was developed to incorporate domain knowledge into the classification process, without excluding specific parts of the landscape or detections therein. Experimental evaluation of WODAN and CarcassonNet showed that these performed reasonable, although there is always room to improve the efficiency and performance of these workflows. The transferability of these methods, albeit with decreased performance, was shown by case studies where both methods were applied on areas with different archaeological, geo(morph)ological and land-use conditions and on LiDAR data with different properties.

In this research, it is argued that the incorporation of automated detection is mostly dependent on the role of these methods in the broader archaeological research framework. This thesis proposes Human—Computer strategies, in which automated detection precedes or is used in conjunction with manual analysis, to highlight areas of archaeological interest that require (field) verification and to add detail to existing archaeological predictive maps. These strategies, with various levels of human involvement depending on the task, resolve many of the current issues of archaeological automated detection methods, while requiring only moderate performance.

Finally, this thesis has shown that the use of automated detection methods can benefit both cultural heritage management and landscape or spatial archaeology (i.e., *räumliche Archäologie*; Doneus, 2013). The efficient detecting and mapping of the presence, location, and distribution of previously unknown archaeological objects within the landscape (i.e., quantitative knowledge gain) is of obvious benefit for both settlement archaeology (*Siedlungsarchäologie*), environmental archaeology (*Umweltarchäologie*) and cultural heritage management. However, such a comprehensive dataset also gives us a more complete view of the patterns and trends within the (large-scale) distribution of archaeological objects in the landscape, the interrelationships between these objects and/or the landscape, and the structuring of the landscape in the past (i.e., qualitative knowledge gain). This forms the basis for a better *understanding* of the archaeological landscape, i.e., landscape archaeology (*Landschaftarchäologie*). Moreover, it offers insight into the current archaeological research practice and possible biases that result from certain methods and/or interpretations.

In conclusion, the application of automated detection methods in large remotely-sensed datasets benefits both cultural heritage management and archaeological research and has the potential to radically transform archaeological practice in the near future.

7.6 Outlook

7.6.1 Combining Methods for Archaeological Automated Detection

A potential benefit for the detection of archaeological objects, especially landscape patterns, lies in combining Deep Learning approaches with other methods that (indirectly) detect archaeological objects and patterns. For instance, combining Deep Learning approaches with predictive modeling (Verhagen et al., 2019) seems an obvious angle (see also Bickler, 2021), especially considering how the use of automated detection results is envisioned in this thesis, i.e., as areas of interest that contain potential archaeological objects. But also combining other techniques (Vletter & van Lanen, 2018) or ontological reasoning (Nuninger et al., 2020a,b) with an automated detection method such as CarcassonNet could lead to an improvement in performance.

7.6.2 Combining Deep Learning and Citizen Science in Archaeology

An interesting angle for further research is the combination of Deep Learning and Citizen Science for archaeological remote sensing. Recently, Citizen Science has been successfully implemented for the detection of archaeological objects in remotely-sensed data (Forest et al., 2020; Lambers et al., 2019; Lin et al., 2014a; Stewart et al., 2020). As discussed by Lambers et al. (2019), Citizen Science might offer solutions to the professional bottleneck apparent in the creation of large training datasets (see Green et al., 2020; Herfort et al., 2019; Keshavan et al., 2018; Willi et al., 2019) and the (field) validation of detections made by automated detection methods. Furthermore, applying Citizen Science and Deep Learning approaches on the same data might offer possibilities to combine results, enhance performance, and alleviate some of the challenges of both methods (Conrad & Hilchey, 2011; Muenich et al., 2016).

7.6.3 Collaboration

The field of Deep Learning moves at an exceptional pace, and original, potentially ground-breaking research is published on a daily basis. This necessitates a considerable time investment in the monitoring of the state-of-the-art, and an ability to properly evaluate recent, often not thoroughly evaluated, research papers and techniques (Olivier & Verschoof-van der Vaart, 2021). Consequently, in archaeological automated detection, generally architectures and methods are 'borrowed' that are proven and currently easily available, but are not necessarily fully understood or state-of-the-art (Cowley et al., 2021). Therefore, for the future development of archaeological automated detection, close collaborations between the field of Archaeology and Computer Science might be very fruitful. Computer scientists have the most up-to-date view of the possibilities of methods, while archaeologists are necessary for defining exact goals of the project, the incorporation of methods and results in digital archaeological practice (e.g., GIS), and for providing domain knowledge (see for instance Fiorucci et al., n.d.). The first competition on the segmentation of Maya structures in LiDAR data (Discover the mysteries of the Maya) seems to have sparked the interest of computer scientists. Hopefully, with the release of archaeological datasets and benchmarks (see below) this interest can result in further collaboration.

7.6.4 Availability of Benchmarks, Datasets, and Methods

In order to further the development of automated detection methods in archaeology, freely available, large, annotated archaeological benchmark datasets are needed (see Opitz & Herrmann, 2018). These could not only offer a performance baseline and equal comparison for existing methods, but could also be used to train and evaluate newly developed methods. Ideally, specific guidelines are drafted, which ensure the compatibility between different datasets through conventions on data formats and annotations. This would offer the opportunity to combine different datasets with a variety of archaeological object classes. The variation in the data would also enhance the generalization capabilities of Deep Learning approaches. In the meantime, research projects should aim to release the developed datasets, although we need to be aware of potential issues with giving up positional data of archaeological objects and looting (see McCoy, 2017).

Finally, next to the availability of datasets, developed techniques and methods (e.g., code, hyperparameters, etc.) should also be made available to enhance future research and to prevent the reinvention of the Deep Learning wheel (Schmidt & Marwick, 2020). Therefore, the datasets and methods created in this research will be made freely available in the near future. Till that time, these are available upon request.