IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 2 June 2023, accepted 23 June 2023, date of publication 28 June 2023, date of current version 6 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3290305

RESEARCH ARTICLE

Uncovering Archaeological Sites in Airborne LiDAR Data With Data-Centric Artificial Intelligence

DANIEL CANEDO[1], JOÃO FONTE[2], LUIS GONÇALVES SECO[2,3], MARTA VÁZQUEZ[4,5], RITA DIAS[6], TIAGO DO PEREIRO[6], JOÃO HIPÓLITO[7], FERNANDO MENÉNDEZ-MARSH[1], PETIA GEORGIEVA[1], (Senior Member, IEEE), AND ANTÓNIO J. R. NEVES[1], (Senior Member, IEEE)

[1]IEETA/DETI, University of Aveiro, 3810-193 Aveiro, Portugal
[2]Department of Communication Sciences and Information Technologies, University of Maia, 4475-690 Maia, Portugal
[3]INESC TEC—Institute for Systems and Computer Engineering, Technology and Science, 4200-465 Porto, Portugal
[4]N2i, Polytechnic Institute of Maia, 4475-690 Maia, Portugal
[5]Center for Research in Geospatial Sciences (CICGE), Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal
[6]ICArEHB, University of Algarve, ERA Arqueologia, 8005-139 Faro, Portugal
[7]ERA Arqueologia, 8005-139 Faro, Portugal

Corresponding author: António J. R. Neves (an@ieee.org)

This work was supported by the Project Odyssey: Platform for Automated Sensing in Archaeology Co-Financed by COMPETE 2020 and Regional Operational Program Lisboa 2020 through Portugal 2020 and FEDER under Grant ALG-01-0247-FEDER-070150.

ABSTRACT Mapping potential archaeological sites using remote sensing and artificial intelligence can be an efficient tool to assist archaeologists during project planning and fieldwork. This paper explores the use of airborne LiDAR data and data-centric artificial intelligence for identifying potential burial mounds. The challenge of exploring the landscape and mapping new archaeological sites, coupled with the difficulty of identifying them through visual analysis of remote sensing data, results in the recurring issue of insufficient annotations. Additionally, the top-down nature of LiDAR data hinders artificial intelligence in its search, as the morphology of archaeological sites blends with the morphology of natural and artificial shapes, leading to a frequent occurrence of false positives. To address this problem, a novel data-centric artificial intelligence approach is proposed, exploring the available data and tools. The LiDAR data is pre-processed into a dataset of 2D digital elevation images, and the known burial mounds are annotated. This dataset is augmented with a copy-paste object embedding based on Location-Based Ranking. This technique uses the Land-Use and Occupation Charter to segment the regions of interest, where burial mounds can be pasted. YOLOv5 is trained on the resulting dataset to propose new burial mounds. These proposals go through a post-processing step, directly using the 3D data acquired by the LiDAR to verify if its 3D shape is similar to the annotated sites. This approach drastically reduced false positives, attaining a 72.53% positive rate, relevant for the ground-truthing phase where archaeologists visit the coordinates of proposed burial mounds to confirm their existence.

INDEX TERMS Archaeology, data-centric artificial intelligence, data augmentation, deep learning, LiDAR, location-based ranking, object detection.

I. INTRODUCTION

Doing fieldwork to search for archaeological sites typically involves covering extensive areas, sometimes with difficult access [1]. Automating this process could streamline many endeavours that expert archaeologists face during this undertaking. Recent advancements in artificial intelligence (AI) have led to significant progress in challenging Computer Vision problems. Using this technology in archaeology to detect archaeological sites could minimize the challenges

The associate editor coordinating the review of this manuscript and approving it for publication was Yizhang Jiang.

associated with labour-intensive and time-consuming fieldwork. However, due to the intricate nature of the images used for this task, Red, Green, and Blue (RGB), multi and hyperspectral, and digital terrain models (DTM), among other types, this remains an unsolved problem. More recently, these tend to be captured using manned or unmanned aerial vehicles (UAV) like drones, resulting in images with a bird's-eye view. This perspective is prevalent in most available data, which is a problem since similar natural and artificial shapes can be mistaken as archaeological sites. Additionally, mapping archaeological sites is difficult, leading to a lack of annotated data in Archaeology, particularly when mapping uncommon archaeological sites that only exist in specific regions.

Recently, there has been a rapid development of UAVs, which is inevitably changing the workflow paradigm in various fields, including Archaeology [2]. Before, archaeologists had to rely mainly on aerial and satellite imagery when conducting surveys. However, this technology has provided archaeologists with a new way to capture their own specialized aerial data. UAVs can be equipped with a Light Detection and Ranging (LiDAR) sensor as it presents several advantages over other sensors or methods in conducting an archaeological mission. LiDAR is capable of producing accurate 3D models of archaeological landscapes, which can then be used for DTM generation [3]. During aerial missions, LiDAR allows the generation of a 3D model of the landscape by calculating the time it takes for the laser to travel from the sensor to being reflected back by the surface. When this laser encounters vegetation, it is partially reflected back to the sensor [4]. This information is recorded and can later be used to remove vegetation when generating the DTMs, providing archaeologists with a detailed topography of the scanned landscape even under dense canopy, potentially revealing archaeological sites that may be hidden under the vegetation. This approach inevitably saves time and reduces costs associated with fieldwork and excavations.

It is possible to apply a panoply of visualization techniques to the DTMs to highlight subtle topographical changes, which generally leads to a better interpretation by both archaeologists and AI [5]. One of those techniques is the Local Relief Model (LRM) [6], which substantially improves the ability to visualize small and shallow topographic features, regardless of the illumination angle, and allows for the direct measurement of their relative elevations and volumes. Figure 1 illustrates this visualization technique applied to a DTM covering the Parque Nacional da Peneda-Gerês, Portugal.

Typically, archaeologists working with LiDAR data take all these steps to achieve a considerably enhanced image without vegetation, which is extremely helpful in their search for new archaeological sites. Although there are other options, such as equipping the drone with a high-resolution camera, they lack the functionalities of LiDAR. Not only it is harder to visualize topographic changes in these visible spectrum images, but it is also not possible to remove the vegetation to see what is beneath it. Furthermore, archaeologists face

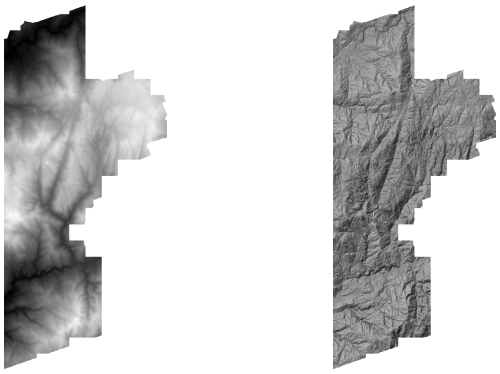

FIGURE 1. An example of DTM and LRM images covering Parque Nacional da Peneda-Gerês, Portugal.

an additional challenge when dealing with visible spectrum images, namely the necessity to perform ground-vegetation segmentation. This arises due to the adverse impact of vegetation on the effectiveness of detecting potential archaeological sites on the ground. Moreover, as concluded in [7], performing the segmentation exclusively on RGB images is a demanding task, since the RGB images need to be complemented with multispectral or thermal images to facilitate the segmentation process. LiDAR sensors, in contrast, circumvent these challenges by providing data that showcases prominent topographic changes and removes the vegetation. This capability makes LiDAR an optimal tool for the detection of archaeological sites, while bypassing the additional complexities and overheads associated with the utilization of a high-resolution camera-based approach. Furthermore, by eliminating the vegetation, previously concealed archaeological sites are revealed and exposed to detection.

The typical pipeline of processing LiDAR data with a visualization technique also shares some problems with conventional imagery. As mentioned previously, since data is generally captured using airborne platforms, it results in a bird's eye perspective which may blend archaeological sites with other similar natural and artificial shapes. Sweeping through this data with a detection algorithm searching for new archaeological sites results in a large number of proposals, but most of them are false positives, as concluded in [8]. Here, the authors successfully reduced the number of false positives by developing a Random Forest classifier [9] to eliminate regions with soils not conducive to the presence of their target archaeological sites: burial mounds. With this approach, they were able to eliminate 324 false positives from 3,086 inferences that were manually validated using the Quantum Geographic Information System (QGIS) software, since they corresponded to rock outcrops, isolated houses, swimming pools, and mound-shaped features of anthropogenic nature. This indicates a 10.5% rate of false positives, which suggests that 9,422 of the 10,527 total inferences across Galicia (Spain) are potential true positives. The authors follow up by stating that this does not mean that 9,422 are burial mounds,

for which proper fieldwork is required to validate those inferences. Additionally, the authors acknowledge that burial mound identifications in forested areas were not considered, as it was impossible for them to identify common occurrences, such as rock outcrops, given their available inspection methods.

The main goal of the research presented in the current manuscript is to identify potential archaeological sites, namely burial mounds, and pinpoint their exact location to the archaeologists. This problem can be solved with either object detection or segmentation. Since burial mounds have a distinct circular shape, object detection is a viable approach. Therefore, in this work we present a novel solution for a typical workflow of archaeological projects using object detection for archaeological mapping. It is proposed a software to break down large images resulting from airborne LiDAR data into a dataset ready to be fed to the most popular object detection algorithms. Additionally, a data augmentation pipeline is proposed based on a copy-paste object embedding and geometric transformations to artificially increase the annotated archaeological sites. The resulting dataset is then used to train the You Only Look Once (YOLO) object detector to infer new burial mounds on unseen data. Finally, the paper concludes by outlining two post-processing validation steps to minimize false positive inferences, namely a Location-Based Ranking (LBR) algorithm and an anomaly detection algorithm.

II. MATERIALS AND METHODS

The region of Alto Minho, Portugal, was considered for the research presented in this paper. The Comunidade Intermunicipal do Alto Minho (CIM Alto Minho) provided us with the airborne LiDAR data from 2018 (2 points per m^2) covering this region (2220 km^2). A visualization technique was applied to the 1-meter LiDAR-derived DTMs, namely the LRM. Four LRM images corresponding to four main sub-regions within Alto Minho were obtained through this process: Viana do Castelo, Paredes de Coura, Arcos de Valdevez, and Parque Nacional da Peneda-Gerês. Table 1 summarizes the characteristics of these images.

TABLE 1. Characterization of the dataset used in this work. For each region it is presented the resolution in pixels, annotations correspond to the number of burial mounds in each image, and the size in gigabytes of the LRM image file.

Region	Resolution (px)	Annotations	Size (GB)
Viana do Castelo	19,978×46,000	14	3.7
Paredes de Coura	13,999×51,999	56	2.9
Arcos de Valdevez	15,999×43,999	71	2.8
Parque Nacional da Peneda-Gerês	19,999×44,955	135	3.6

The chosen object detection algorithm for this work was the fifth version of the popular YOLO object detector [10]. During this work, three new versions of YOLO were released, however, we still kept working with YOLOv5 as it has a stable release from years of iterative development, and it is also integrated with the PyTorch framework. The following subsections present the pipeline of the proposed system.

A. ANNOTATIONS AND PRE-PROCESSING

276 burial mounds were used as training data based on previous research in the Alto Minho region [11], [12]. In order to use these burial mounds to train an object detection algorithm, it was needed to annotate them. QGIS was used for this task and the burial mounds were annotated using polygons. The criterion used to annotate these burial mounds was basically drawing a tight polygon around them. There were two main reasons for using polygons instead of bounding boxes. On one hand, polygons allowed for a precise crop during the copy-paste object embedding used in data augmentation described in Subsection II-B. On the other hand, this annotation approach is future-proof as it can be used by segmentation algorithms directly. In object detection algorithms, they can be used anyway converting polygons into bounding boxes using their extremities. Figure 2 illustrates a cluster of burial mounds in Parque Nacional da Peneda-Gerês annotated with QGIS.

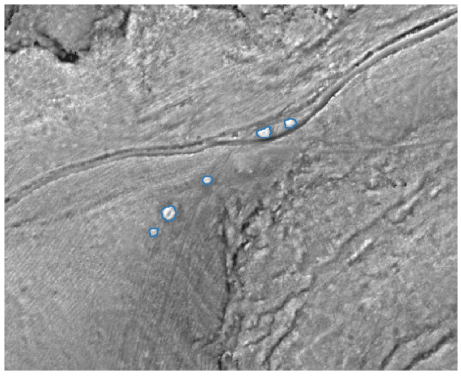

FIGURE 2. Some annotated burial mounds in an LRM image using the QGIS software.

The annotations were saved in Well-Known Text (WKT) format since it is a popular format for representing vector geometry objects, and it is simple to parse. Then, it was needed to process the four LRMs into a usable dataset, since they are essentially four large images with a resolution in the order of tens of thousands of pixels for both rows and columns. Therefore, software was developed to separate these large images into smaller ones with a fixed resolution, constituting a typical YOLO dataset. This process required several steps. Firstly, the LRM and corresponding annotations are loaded. To avoid memory issues, one image is opened at a time. Afterwards, the extent of the image in the Coordinate Reference System (CRS) EPSG:3763 – ETRS89 / Portugal TM06 [13] is parsed from the metadata, and the annotations are converted from polygons to bounding boxes by using their extremities. Then, the bounding boxes are converted from CRS to pixel coordinates and stored into a list. The software iterates over the list to crop images with a fixed resolution around a burial mound. This step is probably the

most important one since it is desired that each burial mound is covered only once in the dataset to maintain uniqueness, and the cropping must be done in a way that every annotated burial mound is completely visible. Therefore, an algorithm was developed to solve this problem by taking a burial mound and creating a range of coordinates around it such that it is possible to generate cropping proposals without compromising the visibility criterion. The following equations were used to generate those coordinates.

$$X_{points} = [X_{max} - \frac{width}{2}, X_{min} + \frac{width}{2}]$$
$$Y_{points} = [Y_{max} - \frac{height}{2}, Y_{min} + \frac{height}{2}] \quad (1)$$

The resolution used for this work is 640 × 640 pixels. Afterwards, these coordinates are shuffled and iterated over. Each coordinate corresponds to the top left corner of the cropping proposal around the burial mound. Then, this cropping proposal is processed to verify if the burial mounds are completely visible and if they do not appear in previously cropped images. If the cropping proposal meets those conditions, the burial mounds are flagged as already being processed such that they do not appear in the next cropped images to keep their uniqueness in the dataset. Figure 3 illustrates the three possible outcomes of this algorithm.

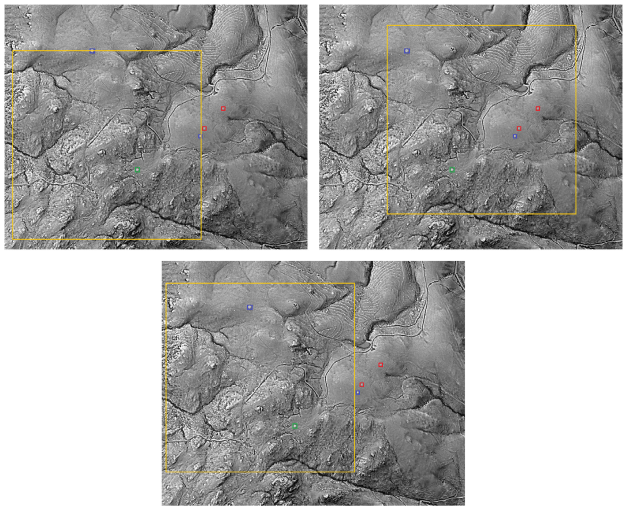

FIGURE 3. Three possible outcomes of the cropping algorithm. The larger yellow bounding box represents the cropping proposal, the green bounding box represents the burial mound that is currently being processed, the blue bounding boxes represent nearby burial mounds, and the red bounding boxes represent burial mounds that were already processed in a previous iteration. On the top left, is a failed iteration because the cropping proposal cuts two burial mounds, not respecting the visibility criterion. On the top right, is a failed iteration because the cropping proposal covers burial mounds from a previous iteration. On the bottom, a successful iteration because the cropping window meets all the criteria.

As it is possible to observe in Figure 3, the larger yellow bounding box represents the cropping proposal, the green bounding box represents the burial mound that is currently being processed, the blue bounding boxes represent nearby burial mounds, and the red bounding boxes represent burial mounds that were already processed in a previous iteration, belonging to another cropped image. The first row of images represents two failed attempts. The top left image failed because the cropping proposal cuts two burial mounds, not meeting the visibility criterion. The top right image failed because the cropping proposal covers burial mounds that already belong to another cropped image from a previous iteration. The second-row image represents a successful attempt. In this image, the cropping proposal meets all the conditions: there are no partially visible burial mounds or burial mounds that already belong to another cropped image. Next, the coordinates of the burial mounds' annotations are mapped into the cropped image and converted to YOLO format. Finally, both cropped images and respective labels are stored, and the next burial mound on the list is processed. At the end of this iterative process, a dataset is generated. Table 2 summarizes this dataset.

TABLE 2. Characteristics of the dataset generated from the available annotations.

Set	Images	Annotations
Training set	146	243
Validation set	18	33

This dataset is in YOLO format and can be directly fed to the YOLO framework. However, during the dataset generation, a few more actions were done in preparation for the next steps of this work. While storing the bounding box labels for the burial mounds, the original polygon annotations were also mapped into the image resolution and stored for later use during data augmentation. Still, during this process, the bounding boxes were used to crop and extract the points corresponding to burial mounds of the raw LiDAR data and store them for later use in Subsection II-D. Figure 4 illustrates the LRM overlaid with the stored points, for better understanding part of the same region of Figure 2 was used.

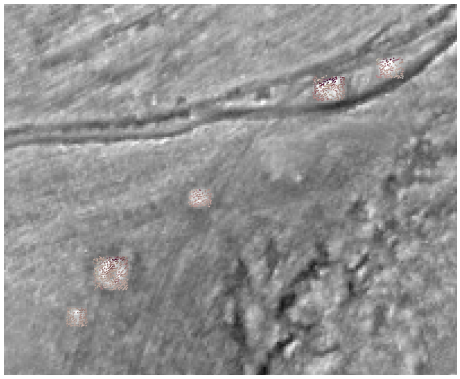

FIGURE 4. LRM data overlaid with point cloud data for each burial mound.

Figure 5 summarizes what was discussed in this Subsection.

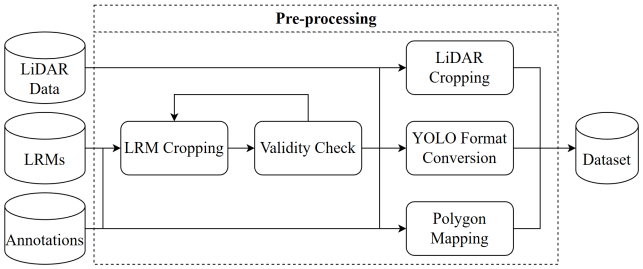

FIGURE 5. Data pre-processing pipeline.

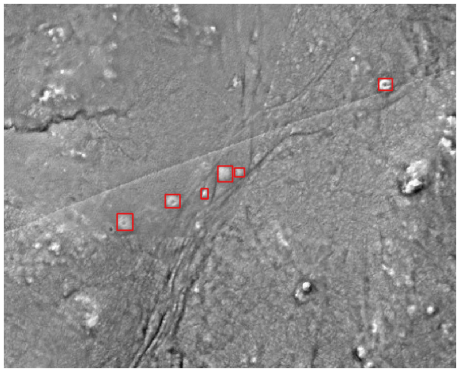

FIGURE 6. Potential cluster of unlabeled burial mounds.

B. DATA AUGMENTATION

It is possible to observe from Table 2 that the dataset size is negligible for such data-hungry Deep Learning algorithms. For instance, YOLOv5 documentation recommends over 1,500 images per class and over 10,000 instances per class [14] for an ideal dataset. However, the available annotated data in this area of research is usually scarce, which results in small datasets. Training on these datasets directly will eventually lead to overfitting [15]. A model overfits when it fits too close to the training data, leading to an accurate classification of that data, at the expense of being unable to accurately predict new data. This phenomenon is obviously undesirable for any classifier, as their main purpose is to classify new data accurately. However, there are ways to minimize this problem, and one of them is data augmentation [16]. The purpose of this technique is to artificially increase and enhance the dataset to satisfy the YOLOv5 requirements for proper training. Typically, this is done through image processing, such as geometric and color transformations. However, there is a lack of background variety in the dataset, which counters one of the recommendations of YOLOv5 [14]. Therefore, this research attempted to solve this problem through a copy-paste object embedding [17], [18].

1) COPY-PASTE DATA AUGMENTATION WITH AUXILIARY YOLOv5

The proposed copy-paste object embedding is a rather complex one, involving several features to ensure robustness. Since one of the main goals of this technique is to increase background variety by pasting burial mounds onto new images, the bulk of the unused data covering Alto Minho when generating the dataset described in Subsection II-A can be used. Only a tiny percentage of the data was used because the images constituting the dataset were cropped around annotated burial mounds. Therefore, new images can be cropped from unused data, increasing the background variety naturally. However, a problem arises from this approach. As fieldwork was not carried out for most of the Alto Minho region, there are certainly unidentified tombs, which means that new images of unused data cannot be randomly cropped to insert new instances. If this technique crops a new image that contains an unlabeled burial mound, this image will deteriorate the YOLOv5 training results, as it has a contradicting instance instructing the model that it does not belong to the burial mound class.

Figure 6 shows a hypothetical new image cropped from unused data to better illustrate this aforementioned problem. This image was analyzed by archaeologists with expertise in remote sensing, and they reached the consensus that the outlined instances are most likely burial mounds. However, they were not verified on the field and, consequentially, they were not part of the original 276 annotations. Therefore, this image has potentially unlabeled burial mounds, breaking the label consistency required for a proper training, and it should not be used for the copy-paste object embedding.

The data augmentation pipeline should be equipped with a tool to identify potential burial mounds, which in a way is a paradox, as this is exactly the goal of the AI model. However, the purpose of such identification is to discard the proposed cropped images as they potentially contain unlabeled burial mounds. The solution found for this problem was training a YOLOv5 model (see Subsection II-C for details) with the original dataset generated in Subsection II-A. This model is responsible for identifying potential burial mounds in the new cropped images, such that they can be discarded from the data augmentation process. Since this model was trained with little data, it is highly sensible and detects every shape that has any resemblance to a burial mound, resulting in many inferences, mostly false positives. Despite the high false positive rate, this model is still useful as it ensures that all potential burial mounds are covered. And since the unused data is quite large, the background variety is not compromised, even if there are many discarded cropped images.

Another problem that arises from the copy-paste object embedding is how to choose the pasting location. If the pasting is done randomly, the pasted burial mounds could fall over rivers, houses, streets, and so on. That is not consistent with the real-world, as burial mounds only exist in certain locations. For this reason, a new approach was developed, inspired by the LBR technique proposed by [19]. In this proposal, the authors ranked the most probable locations to find barrows in Veluwe (Netherlands), following their deduction that the location of archaeological sites in the landscape is not random but is the result of certain characteristics of the past and

present environment. They continue by elaborating that those characteristics are mainly subsoil and land-use. Following their work and conclusions, this research attempts to propose a novel approach where the Land-Use and Occupation Charter of Portugal [20] is used. This charter is essentially a raster image covering Portugal, with 83 classes obtained through visual interpretation. The 2018 version was chosen to match the coverage year of the provided LiDAR data for Alto Minho.

Figure 7 illustrates the Land-Use and Occupation Charter of Portugal from 2018. Each color represents different classes: forests, lakes, rivers, streets, buildings, infrastructures, and agriculture, just to name a few. Basically, the idea for using this charter in data augmentation is to outline the most probable regions to find burial mounds, such that during the copy-paste object embedding, burial mounds are pasted onto coherent regions. To find these regions, the 276 validated burial mounds were used. Essentially, the classes overlapping with these burial mounds were considered as probable locations. These locations revolve around forests, sparse vegetation, and agriculture. All the other locations were removed, such as buildings, streets, rivers, and infrastructures. With the regions of interest properly outlined, the pasting of burial mounds onto new images can be done more freely.

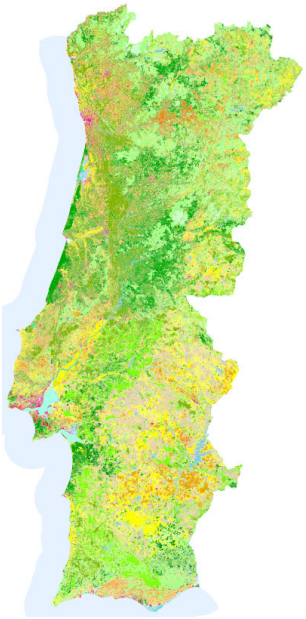

FIGURE 7. Land-Use and Occupation Charter of Portugal, 2018. Shades of green represent different types of forests, shades of yellow represent different types of agriculture, shades of red represent infrastructures.

Since the two main problems were overcome, the data augmentation pipeline could be built. As it was done in Subsection II-A, one LRM is loaded at a time to avoid memory issues. The original polygon annotations described in Subsection II-A for that LRM are loaded. The reason for using the original polygons instead of bounding boxes is because the cropping is guaranteed to only contain the burial mounds and not any background that results from rectangular cropping. With this, the problem of YOLOv5 learning to detect rectangular disturbances instead of actually learning to detect burial mounds is avoided. Then, a random crop within a region of interest is done to the LRM. The YOLOv5 model described previously attempts to infer any potential burial mounds in this cropped image. If there are any potential burial mounds, the image is discarded and another crop is done. If the image is free from potential burial mounds, the process continues. Next, a number of burial mounds are randomly chosen to be pasted onto this new image. Coordinates for each pasting are randomly generated following two main requirements: they must not overlap, and they must be within the regions of interest. Figure 8 illustrates a hypothetical cropped image.

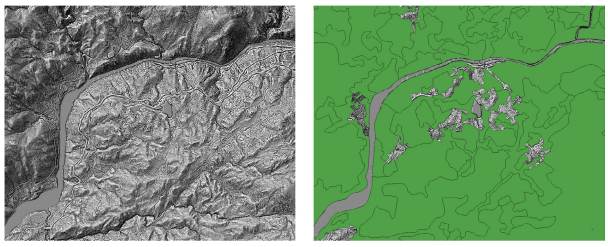

FIGURE 8. Cropped image from LRM on the left, and the same image overlaid with the proposed LBR on the right.

As it is possible to observe in the right image of Figure 8, the green regions are the ones of interest that resulted from the Land-Use and Occupation Charter of Portugal mentioned previously. In this example, there is a river and some small regions that were discarded during the filtering process. Here, the burial mounds would only be pasted onto the green regions without overlapping. Before the pasting is done, the burial mounds are smoothed to eliminate any sharp edges resulting from their cropping, and a random geometric transformation is applied to both the burial mounds and the cropped image.

All of this is done in an iterative manner, for as many iterations as desired to artificially generate new data. It is also worth mentioning that this pipeline is prepared to respect class balance, although it was not needed for this work since only one class was tackled. Additionally, this process generates around 10% of background images without any burial mound, respecting the YOLOv5 recommendations [14]. Figure 9 summarizes what was discussed in this Subsection.

Using the proposed data augmentation pipeline, the dataset obtained in Subsection II-A and presented in Table 2 was augmented. Table 3 summarizes the copy-paste augmentation dataset.

Figure 10 shows two samples resulting from the copy-paste data augmentation.

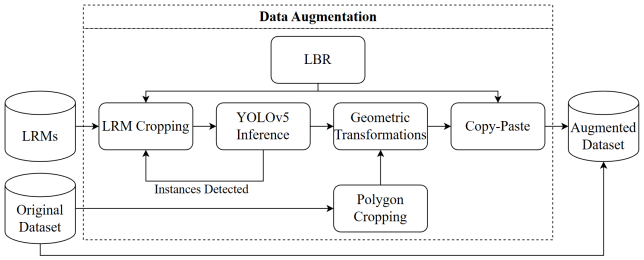

FIGURE 9. Data augmentation pipeline.

TABLE 3. Characteristics of the copy-paste augmentation dataset.

Set	Images	Annotations
Training set	2482	21248
Validation set	306	2598

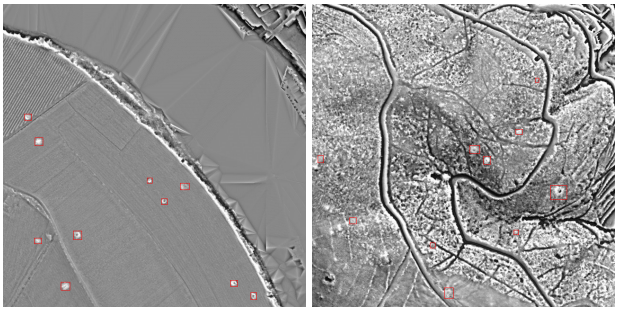

FIGURE 10. New images resulting from the copy-paste data augmentation.

2) TRADITIONAL DATA AUGMENTATION

Just to compare how a traditional data augmentation pipeline fairs against the proposed copy-paste data augmentation, another dataset was augmented from the original one generated in Subsection II-A, using the Albumentations software [21]. Here, the augmentation pipeline follows a traditional structure, where images are randomly flipped, and a random color transformation is applied from the following ones: color jitter, saturation value, gamma value, RGB shift, brightness, and contrast. This color transformation was included here to avoid having equal samples in the dataset. Table 4 summarizes the traditional augmentation dataset.

TABLE 4. Characteristics of the traditional augmentation dataset.

Set	Images	Annotations
Training set	2482	4131
Validation set	306	561

Comparing Table 3 with Table 4, it is possible to observe one main advantage of the proposed copy-paste data augmentation over the traditional one. In the copy-paste data augmentation, the number of annotations can be freely increased independently on the number of images, while in the traditional data augmentation, the increase of annotations is tied to the increase of their belonging images.

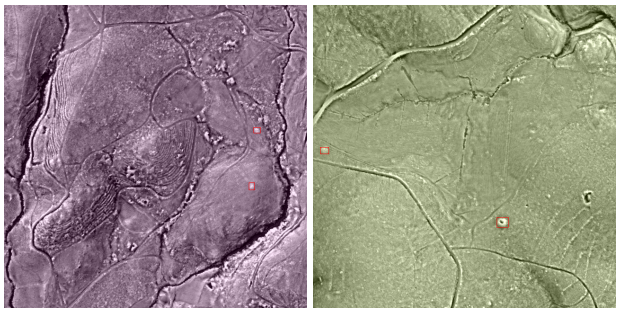

FIGURE 11. New images resulting from the traditional data augmentation.

Additionally, the copy-paste data augmentation increases the background variety. Figure 11 shows two samples resulting from the traditional data augmentation. Note that in this procedure there was no copy-paste, as the burial mounds seen in the images are the original ones.

C. TRAINING

As mentioned previously, YOLOv5 was the chosen object detection algorithm for this work. Its architecture consists of a Convolutional Neural Network (CNN) with three main components: backbone, neck, and head. YOLOv5 integrates the Cross Stage Partial Network (CSPNet) [22] combined with Darknet as its backbone to extract features. This backbone proves advantageous by addressing redundant gradient information, reducing the model's size, and enhancing both accuracy and speed when compared to previous iterations of the YOLO framework. For the neck, YOLOv5 employs the Path Aggregation Network (PANet) [23], which effectively augments the localization capability by propagating strong responses from low-level features. This is based on the understanding that edges or instance parts serve as robust indicators for accurately localizing objects. Finally, the head of YOLOv5 enables multi-scale predictions, further contributing to its overall effectiveness.

YOLOv5 offers a panoply of networks that differ in size. The medium network YOLOv5m6 was chosen for this work based on a compromise between processing time and accuracy. Based on the YOLOv5 benchmark on the COCO dataset [24], YOLOv5m6 scores a mean Average Precision of 69.3 calculated with an intersection over union threshold of 0.5 (mAP@0.5), while the largest network (YOLOv5 $\times$ 6) scores a mAP@0.5 of 72.7. This is just a slight improvement of 4.9%, although YOLOv5 $\times$ 6 is 294.1% larger than YOLOv5m6. Table 5 illustrates this comparison.

TABLE 5. YOLOv5 parameters and respective mAP on the COCO dataset for each network. On the column labels, the "6" refers to the YOLOv5-P6 models, which have 4 output layers at strides 8, 16, 32, and 64. From left to right: nano (n), small (s), medium (m), large (l), extra large (x).

	n6	s6	m6	l6	x6
Parameters (Millions)	3.2	12.6	35.7	76.8	140.7
mAP@0.5	54.4	63.7	69.3	71.3	72.7

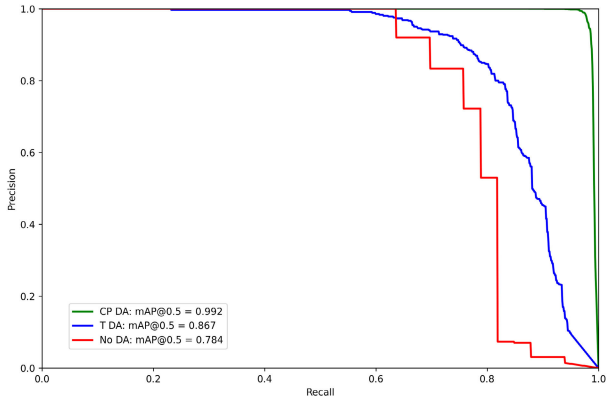

FIGURE 12. Precision-recall curves on the validation set. The red curve represents the YOLOv5m6 model trained on the original dataset with no data augmentation (No DA); the blue curve represents the YOLOv5m6 model trained on the traditional augmentation dataset (T DA); the green curve represents the YOLOv5m6 model trained on the copy-paste augmentation dataset (CP DA).

In order to study the impact of the proposed copy-paste data augmentation, a YOLOv5m6 model was trained on the original dataset (see Table 2), another was trained on the copy-paste augmentation dataset (see Table 3), and another was trained on the traditional augmentation dataset (see Table 4). Transfer Learning [25] was implemented, starting the training with the pre-trained weights on the COCO dataset. The used optimizer was Stochastic Gradient Descent (SGD) [26] with a decay of 0.0005 and a learning rate of 0.01. The image resolution was set to 640 × 640, the batch size was set to 15, and the models were trained for 300 epochs, stopping earlier if training results did not improve for 20 consecutive epochs. The training was done with an Nvidia GeForce RTX 3080 10 GB GDDR6X GPU and an AMD Ryzen 5 5600X 6-Core 3.7GHz CPU. Figure 12 illustrates the precision-recall curves on the validation set.

D. INFERENCE AND POST-PROCESSING VALIDATION

In this Section, the proposed inference pipeline and two post-processing validation algorithms are described. The inference process is simple: a sliding window with a step of half of the image resolution sweeps through the testing LRM images. Although this window step will inevitably generate more inferences, it guarantees that each burial mound is completely visible at least once. Since burial mounds will most likely be detected more than once because of this window step, non-maximum suppression is applied to deal with overlapping inferences. From each window, an image with a 640 × 640 resolution is cropped and fed into the YOLOv5m6 model to detect burial mounds. Afterwards, these inferences go through two post-processing validation algorithms. These algorithms have the goal of filtering as many false positives as possible during the inference process. As mentioned in Section I, false positives are a common problem when attempting to detect archaeological sites in airborne LiDAR data, especially because of the bird's-eye perspective. Sweeping through this data that typically cover large regions with an object detection algorithm will inevitably lead to a considerable number of false positives with the current technology, as these archaeological sites are often mistaken with other similar shapes.

The first validation algorithm is the LBR discussed in Subsection II-B for the copy-paste data augmentation. Since LBR allows for the outlining of probable regions to find burial mounds, this validation step simply verifies if the inferences are located in those probable regions (see right image of Figure 8). If they are found within a probable region, they are passed on to the second validation algorithm. If they are not found within a probable region, they are most likely false positives and therefore discarded.

The second validation algorithm is the Local Outlier Factor (LOF) [27]. This algorithm is an unsupervised anomaly detection method that assesses the density deviation of a data point compared to its neighbours. If a data point has a considerably lower density, this algorithm classifies it as an anomaly. The idea for this validation step is to use the 3D information that is lost when converting the LiDAR data to the LRM. The hypothesis is that this 3D information can be crucial in distinguishing between a burial mound and a false positive that happens to have a similar shape in a bird's-eye view.

When generating the dataset with the original annotations described in Subsection II-A, points associated with burial mounds were cropped out of the LiDAR point cloud data and stored in preparation for this validation step (see Figure 4). For feature extraction, a Python implementation of the CloudCompare software [28], [29] was used. With this, it was possible to extract 14 features from each burial mound point, namely Sum of Eigenvalues, Omnivariance, Eigenentropy, Anisotropy, Planarity, Linearity, Principal Components (PC1, PC2), Surface Variation, Sphericity, Verticality, and Normal Vector (Nx, Ny, Nz). Since each burial mound point cloud data has dozens or hundreds of points and each point contributes to 14 features, the resulting number of features for the whole dataset can be massive. For this reason, the median, variance, standard deviation, and covariance were calculated for each burial mound point cloud data. This means that each burial mound contributes with 56 features (14 × 4), which is a significant feature reduction, while maintaining the overall information. An LOF algorithm was trained on 276 point clouds of burial mounds corresponding to the burial mounds from the training and validation sets of the original dataset (see Table 2). Figure 13 summarizes what was discussed in this Subsection.

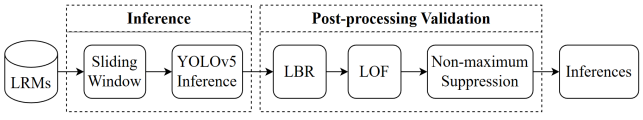

FIGURE 13. Inference and post-processing validation pipeline.

III. RESULTS AND DISCUSSION

As described in Subsection II-D, a sliding window went through the whole region of Alto Minho, feeding the trained

TABLE 6. YOLOv5m6 inferences for the Alto Minho region. From left to right: No data augmentation (No DA), traditional data augmentation (TDA), copy-paste data augmentation (CP DA), copy-paste data augmentation with LBR (CP DA + LBR), copy-paste data augmentation with LBR and LOF (CP DA + LBR + LOF).

	No DA	T DA	CP DA	CP DA + LBR	CP DA + LBR + LOF
Inferences	3417	1558	1467	1124	**648**

models with cropped images. Any inference of known burial mounds is discarded, as they were used on the training step. Table 6 shows the obtained results for this experiment.

From Table 6, *No DA* refers to the YOLOv5m6 model trained with the original dataset, *T DA* refers to the YOLOv5m6 model trained with the Traditional data augmentation dataset, and *CP DA* refers to the YOLOv5m6 model trained with the copy-paste data augmentation dataset (see Subsection II-C for training details).

The goal of this experiment was to select the most robust model, considering the one which produced the least amount of inferences. For this reason, the YOLOv5m6 model trained on the *CP DA* dataset was selected, with 1467 inferences. These inferences went through the post-processing validation described in Subsection II-D. The first step (LBR) was responsible for removing inferences within less probable regions, shortening their number to 1124 inferences. The second step (LOF) was responsible to validate inferences based on their 3D information similarity to burial mounds, further reducing the number to 648 inferences. Then, four expert archaeologists proceeded with a digital validation of these inferences. For this digital validation, they used QGIS and were aided by the LiDAR-derived LRMs, Google Satellite images, Bing Aerial images, and aerial images from Direção-Geral do Território (DGT), which is the Portuguese territory institution, dating from 2021, 2018, 2004-2006, and 1995. Table 7 shows the results from the digital validation.

TABLE 7. Digital validation by four archaeologists of the 648 inferences (see Table 6) produced by the proposed approach.

Potential True Positives	False Positives	Positive Rate
470	178	72.53%

According to the digital validation, 470 inferences are potential true positives while 178 inferences are false positives, which indicates a 72.53% positive rate. Table 8 shows the interpretation of the archaeologists regarding false positives.

TABLE 8. False positives.

Roundabouts	Buildings	Rock Outcrops	Others
3	15	129	31

Most inferences in the *Others* column are false positives that the archaeologists could not interpret, but it also includes construction sites and structures. As it can be observed in Table 8, the bulk of false positives are rock outcrops, which was expected since [8] reported the same problem. What is more interesting are the roundabouts, buildings, construction sites, and structures inferences. These inferences should have been filtered by the LBR validation step, but they were not because of an inconsistency: some roads and infrastructures are not fully outlined in the charter.

Figure 14 illustrates the referred problem. The red color represents buildings and roads, the yellow color represents different types of agriculture, and the green color represents different types of forests. In the three columns, the red segmentation does not seem to properly outline the road. In the first and second columns, it can be observed that roundabouts were mistaken as burial mounds because the LBR charter is inconsistent with the real world. In the second column, the house connected to the road is also not properly outlined by the red color. This fact places the inferences within an agricultural region, which is considered a region of interest. In the third column, it is possible to observe a small infrastructure in DGT 2018 that was not reflected in the LBR charter, resulting in the inference being placed within an agriculture region, which is a region of interest. These limitations with the Land-Use and Occupation Charter of Portugal were expected due to its equivalent scale of 1:25,000. This scale may not accurately represent isolated features of smaller scale. However, this validation step allowed for the removal of 343 false positives, which is a major help in tackling this problem.

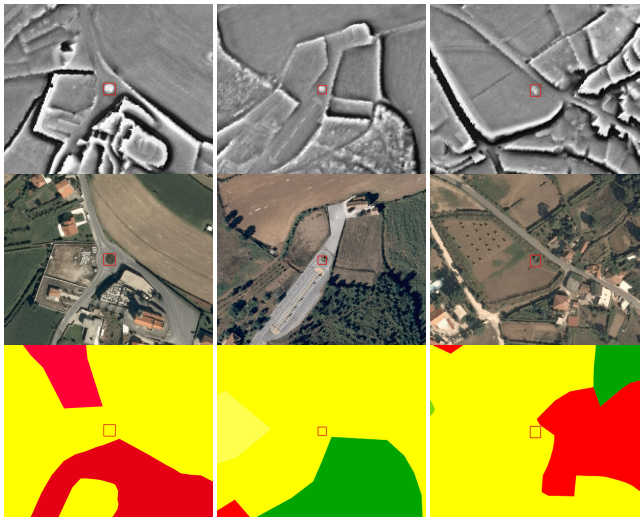

FIGURE 14. Inconsistent LBR charter. From top to bottom: LRM (2018), DGT 2018, and LBR (2018). The first and second columns represent a roundabout inference, and the third column represents a small infrastructure inference. In the last row, it is possible to observe these inferences being wrongly placed within a probable region (agriculture), which validated them.

In the second validation step (LOF) it is attempted to remove false positives based on their 3D LiDAR information. This step allowed for the removal of 476 false positives,

FIGURE 15. Some examples of false positives due to rock outcrops. On the left, the LRM images. On the right, the corresponding Google Satellite images.

FIGURE 16. Example of potential true positives. On the left, the LRM images. On the right, the corresponding Google Satellite images.

which is a significant reduction. However, structures with an overall similar 3D shape to burial mounds were always validated as burial mounds. This was the case for rock outcrops, as seen in Table 8. The Alto Minho region is full of rock outcrops, and all of them are placed within regions of interest, mainly forests. This got them validated by the first validation step. Consequently, similar 3D shapes to burial mounds go through the second validation step and are also validated.

As it is possible to observe in the LRM image presented in Figure 15, the morphology of the detected sites is quite similar to that of burial mounds. Firstly, their similar morphology induced the YOLOv5m6 model in error. Secondly, the fact that they are within a region of interest (forest) got them validated by the first validation step (LBR). Finally, their 3D LiDAR information is similar to that of burial mounds, which got them validated by the second validation step (LOF). This is the most common type of false positive, as it can be observed in Table 8. For future work, this problem will be tackled by adding one last validation step. In this validation step, multispectral imagery will be used to train a classifier with two classes: burial mounds and false positives. It is expected that with this method, false positives like rock outcrops can be properly filtered.

Nonetheless, the proposed approach attained 72.53% positive rate. Figure 16 illustrates some potential true positives.

IV. CONCLUSION

The proposed data-centric AI approach presented in this paper shows great potential for assisting archaeologists in identifying potential archaeological sites. The approach includes a novel data processing method, data augmentation technique, training and inference process, and post-processing step, which together significantly reduce false positives in identifying burial mounds. The positive rate achieved in the Alto Minho region was 72.53%, demonstrating the effectiveness of the proposed approach.

Firstly, a method to break large chunks of data into a dataset formatted to the YOLO framework is discussed. Secondly, a data augmentation method based on copy-pasting known burial mounds to other unused regions is presented. Thirdly, the training step and inference process of YOLOv5 is outlined. Finally, a post-processing step based on two validation steps is proposed. The first validation step verifies if the YOLOv5 inferences are within a region of interest, and the second validation step verifies if those inferences have a similar 3D shape to the burial mounds.

Some issues were identified during the experimental work described in this paper, such as inconsistencies in the LBR charter used in the first validation step and the presence of a large number of rock outcrops in the Alto Minho region that passed the second validation step since they had a similar 3D shape when comparing to the burial mounds.

The goal for future work is the introduction of another validation step in which multispectral images will be used to help distinguish between the surface of burial mounds and the surface of false positives such as rock outcrops.

ACKNOWLEDGMENT

The authors would like to thank the Comunidade Intermunicipal do Alto Minho (CIM Alto Minho) for providing the airborne LiDAR data used in this work.

REFERENCES

[1] P. Drewett, *Field Archaeology: An Introduction*. London, U.K.: Routledge, 2011.
[2] O. Risbøl and L. Gustavsen, "LiDAR from drones employed for mapping archaeology—Potential, benefits and challenges," *Archaeol. Prospection*, vol. 25, no. 4, pp. 329–338, Oct. 2018.
[3] S. I. Jiménez-Jiménez, W. Ojeda-Bustamante, M. Marcial-Pablo, and J. Enciso, "Digital terrain models generated with low-cost UAV photogrammetry: Methodology and accuracy," *ISPRS Int. J. Geo-Inf.*, vol. 10, no. 5, p. 285, Apr. 2021.
[4] F. Remondino and S. Campana, "3D recording and modelling in archaeology and cultural heritage," *BAR Int. Ser.*, vol. 2598, pp. 111–127, Jan. 2014.
[5] A. Guyot, M. Lennon, and L. Hubert-Moy, "Objective comparison of relief visualization techniques with deep CNN for archaeology," *J. Archaeol. Sci., Rep.*, vol. 38, Aug. 2021, Art. no. 103027.
[6] R. Hesse, "Using LiDAR-derived local relief models (LRM) as a new tool for archaeological prospection," in *Landscape Archaeology Between Art and Science*. Amsterdam, The Netherlands: Amsterdam University Press, 2015, p. 369.
[7] H. A. Orengo and A. Garcia-Molsosa, "A brave new world for archaeological survey: Automated machine learning-based potsherd detection using high-resolution drone imagery," *J. Archaeol. Sci.*, vol. 112, Dec. 2019, Art. no. 105013.
[8] I. Berganzo-Besga, H. A. Orengo, F. Lumbreras, M. Carrero-Pazos, J. Fonte, and B. Vilas-Estévez, "Hybrid MSRM-based deep learning and multitemporal Sentinel 2-based machine learning algorithm detects near 10k archaeological tumuli in North-Western Iberia," *Remote Sens.*, vol. 13, no. 20, p. 4181, Oct. 2021.
[9] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," *ISPRS J. Photogramm. Remote Sens.*, vol. 114, pp. 24–31, Apr. 2016.
[10] G. Jocher et al., "Ultralytics/YOLOv5: V5.0—YOLOv5-P6 1280 models, AWS, supervise.ly and YouTube integrations," *Zenodo*, 2021. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2021zndo...4679653J/abstract
[11] M. G. Sousa. (2013). *O Fenómeno Tumular e Megalítico da Região Galaico-Portuguesa do Miño*. [Online]. Available: http://hdl.handle.net/10347/7510
[12] A. M. S. Bettencourt and L. V. Boas, "Monumentos megalíticos do Alto Minho. Uma paisagem milenar," in *Viagem no Tempo. História e Património Cultural do Alto Minho*, A. Campelo, Ed. Viana do Castelo, Portugal: Comunidade Intermunicipal do Alto Minho, 2021, pp. 35–44.
[13] *PT-TM06/ETRS89, Portugal Reference Systems*. Accessed: May 8, 2023. [Online]. Available: https://www.dgterritorio.gov.pt/geodesia/sistemas-referencia/portugal-continental/PT-TM06-ETRS89?language=en
[14] G. Jocher. *Tips for Best Training Results*. Accessed: Mar. 23, 2023. [Online]. Available: https://github.com/ultralytics/yolov5/wiki/Tips-for-Best-Training-Results
[15] D. M. Hawkins, "The problem of overfitting," *J. Chem. Inf. Comput. Sci.*, vol. 44, no. 1, pp. 1–12, Jan. 2004.
[16] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," 2017, *arXiv:1712.04621*.
[17] G. Ghiasi, Y. Cui, A. Srinivas, R. Qian, T. Lin, E. D. Cubuk, Q. V. Le, and B. Zoph, "Simple copy-paste is a strong data augmentation method for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2918–2928.
[18] D. Dwibedi, I. Misra, and M. Hebert, "Cut, paste and learn: Surprisingly easy synthesis for instance detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1301–1310.
[19] W. B. Verschoof-van der Vaart, K. Lambers, W. Kowalczyk, and Q. P. J. Bourgeois, "Combining deep learning and location-based ranking for large-scale archaeological prospection of LiDAR data from The Netherlands," *ISPRS Int. J. Geo-Inf.*, vol. 9, no. 5, p. 293, May 2020.
[20] DGT. *Cartografia de Uso e Ocupação do Solo*. Accessed: Mar. 23, 2023. [Online]. Available: https://smos.dgterritorio.gov.pt/cartografia-de-uso-e-ocupacao-do-solo
[21] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and flexible image augmentations," *Information*, vol. 11, no. 2, p. 125, Feb. 2020.
[22] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 390–391.
[23] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8759–8768.
[24] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Computer Vision—ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham, Switzerland: Springer, 2014, pp. 740–755.
[25] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2009.
[26] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. COMPSTAT*, Y. Lechevallier and G. Saporta, Eds. Paris, France: Physica-Verlag, 2010, pp. 177–186.
[27] O. Alghushairy, R. Alsini, T. Soule, and X. Ma, "A review of local outlier factor algorithms for outlier detection in big data streams," *Big Data Cogn. Comput.*, vol. 5, no. 1, p. 1, Dec. 2020.
[28] Jakteristics. *Python Package to Compute Point Cloud Geometric Features*. Accessed: Mar. 27, 2023. [Online]. Available: https://jakteristics.readthedocs.io/en/latest/
[29] CloudCompare. *3D Point Cloud and Mesh Processing Software*. Accessed: Mar. 27, 2023. [Online]. Available: https://www.danielgm.net/cc/

DANIEL CANEDO received the M.S. degree in electronics and telecommunications engineering from the University of Aveiro, Portugal, in 2017, where he is currently pursuing the Ph.D. degree in computer engineering.

Since 2017, he has been a Research Fellow with the Institute of Electronics and Informatics Engineering of Aveiro. He has published in several international conference proceedings and journals. His research interests include computer vision and artificial intelligence.

Mr. Canedo's awards and honors include the first place on the NATO StratCom Competition—"How to detect malicious use of video and/or photographic content online?," Riga, Latvia, in December 2018.

JOÃO FONTE received the Ph.D. degree in archaeology from the University of Santiago de Compostela, Spain, in 2015.

Since then, he has held several postdoctoral positions in different academic and scientific institutions with the University of Exeter, U.K., Universidade da Maia, Portugal, and Incipit, CSIC, Spain. He is an archaeologist with expertise in archaeological remote sensing, GIS, machine learning, and landscape archaeology. He has published extensively on these topics in several high-impact international journals.

LUIS GONÇALVES SECO received the Ph.D. degree in computer science from the University of Santiago de Compostela, Spain, in 2007.

He is currently an Associate Professor with the Department of Communication Sciences and Information Technologies, University of Maia, Portugal. Besides this, he has been conducting research in research and development units of the National Scientific and Technological System. He is also a Senior Researcher with the INESC TEC—Institute for Systems and Computer Engineering, Technology and Science. His research interests include geospatial sciences, namely what concerns LiDAR technology and multispectral imaging. He has published numerous articles in international journals.

MARTA VÁZQUEZ received the B.S. and Ph.D. degrees in chemical engineering from the Faculty of Engineering, University of Porto, Portugal, in 1994 and 2002, respectively.

She is currently an Assistant Professor with the Polytechnic Institute of Maia. Besides this, she has been conducting research in research and development units of the National Scientific and Technological System. She is also a Researcher with the Center for Research in Geospatial Sciences (CICGE), Faculty of Sciences, University of Porto. She has published more than 30 articles in international journals. Her research interests include numerical analysis, simulation and optimization, statistical analysis, and experimental data processing applied to different scientific fields.

RITA DIAS received the Ph.D. degree in archaeology from the University of Algarve, Portugal, in 2018.

She is currently a Project Manager with ERA Arqueologia and an integrated Researcher with ICArEHB. She is also an Archaeologist, specialized in zooarchaeology and prehistory, more specifically in ichthyological remains and seasonality studies through the use of sclerochronology and stable isotopes analysis. Lately, she has been involved in remote sensing research through her participation as the Project Leader of the Odyssey Remote Sensing Project and also as a UAV Pilot. She has published in *Archaeology International* peer reviewed journals and has organized and participated in several conferences, proceedings, and other publications.

TIAGO DO PEREIRO received the master's degree in geoarchaeology from the Faculty of Sciences, University of Lisbon, in 2014.

He is currently the Director of the Archaeological Survey Department, ERA Arqueologia, carrying out geophysical prospecting and remote sensing work.

JOÃO HIPÓLITO received the degree in archaeology from the University of Porto, in 2009. He joined the Era Arqueologia, in 2008, where he incorporated teams in all types of archaeological interventions. With particular interest in the more technical and practical variant of archaeological study, he focused on new technologies applied to graphic recording, to which he has dedicated his career and as a UAV pilot. He received the Archaeological Assistant Course from the Professional School of Archaeology, Marco de Canaveses, in 2004, and specialized in archaeology technical drawing, immediately starting his professional activity in IPPAR.

FERNANDO MENÉNDEZ-MARSH received the B.A. degree in archaeology and the M.S. degree in geographic information technologies from the Complutense University of Madrid, in 2020 and 2021, respectively.

He is currently a Predoctoral Researcher with the Institute of Electronics and Informatics Engineering, University of Aveiro, Portugal. His research interests include landscape archaeology and how to effectively use geospatial technologies for archaeological purposes.

PETIA GEORGIEVA (Senior Member, IEEE) received the master's and Ph.D. degrees in control system engineering from the Technical University of Sofia (TUS), and the Ph.D. degree in computer science from the Faculty of Engineering of University of Porto (FEUP), Portugal, in 2003. She is currently a Professor in machine learning with the Department of Electronics Telecommunications and Informatics (DETI), University of Aveiro, Portugal, and a Senior Researcher with the Institute of Electronics Engineering and Telematics of Aveiro (IEETA), Portugal. Her research interests include machine learning, deep learning and data mining with strong application focus on image processing, autonomous driving, wireless communications, and brain computer interfaces. She is a Senior Member of the International Neural Network Society (INNS). Her research is funded by sponsors, such as EU, Portuguese Foundation for Science and Technology (FCT), and industry.

ANTÓNIO J. R. NEVES (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the University of Aveiro, Portugal, in 2007.

He is currently an Assistant Professor with the Department of Electronics, Telecommunications, and Informatics Engineering, University of Aveiro. Since 2002, he has been a Research Fellow with the Institute of Electronics and Informatics Engineering of Aveiro. He is the author of numerous research papers published in international conference proceedings, books, and international journals. He has participated in more than 30 research projects involving computer vision, signal processing, robotics, and intelligent systems, some of them still ongoing. As a result of his research, he received more than 35 awards. His research interests include image processing, computer vision, image compression, artificial intelligence, intelligent systems, health information systems, mobile and intelligent robotics, and robotics in industry.