Journal of Cultural Heritage 52 (2021) 171–183

Contents lists available at ScienceDirect

Journal of Cultural Heritage

journal homepage: www.elsevier.com/locate/culher

Original article

Deep learning to detect built cultural heritage from satellite imagery. - Spatial distribution and size of vernacular houses in Sumba, Indonesia -

Fabrice Monna[a,*], Tanguy Rolland[a], Anthony Denaire[a], Nicolas Navarro[b,c], Ludovic Granjon[d], Rémi Barbé[d], Carmela Chateau-Smith[e]

[a] ARTEHIS, UMR CNRS 6298, Université de Bourgogne–Franche Comté, 6 boulevard Gabriel, Bât. Gabriel, F-21000 Dijon, France
[b] EPHE, PSL Research University, F-21000 Dijon, France
[c] Biogéosciences UMR CNRS 6282, Université Bourgogne Franche-Comté, 6, boulevard Gabriel, Bat. Gabriel, F-21000 Dijon, France
[d] MSH de Dijon, USR CNRS 3516, Université Bourgogne Franche-Comté, 6, esplanade Erasme, F-21000 Dijon, France
[e] CPTC, Université de Bourgogne, 4, boulevard Gabriel, F-21000 Dijon, France

ARTICLE INFO

Article history:
Received 6 May 2021
Accepted 12 October 2021
Available online 31 October 2021

Keywords:
Object detection
Data augmentation
Spatial distribution
Settlement
Geographical information system
Scan statistics
Histogram of gradients

ABSTRACT

In Sumba Island – Indonesia, the implantation of vernacular houses, inside and outside traditional villages, is considered to be an efficient proxy for the on-going complex cultural transformations resulting from globalization. This study presents an easily reproducible workflow allowing buildings to be automatically detected from satellite imagery, demonstrating how modern computer vision methods based on deep learning can help in this task, which would be far too time-consuming when undertaken by hand. Eight deep learning architectures based on convolutional neural networks were compared in terms of ability to identify and locate precisely traditional houses from satellite images. By combining a Faster R-CNN ResNet 101 architecture with artificial data augmentation, the model was taught properly using 1033 instances of vernacular houses (AP@.50:.95 = 71.9). Once 14 952 traditional houses had been detected, the Histogram of Orientated Gradients (HOG) was computed and processed by several machine learning algorithms to assess their surface area, as this parameter conveys pertinent information about the economic and political position of the householder. The best classifier was found to be a support vector regressor (SVR, $R^2 = 0.88$), although the other classifiers tested also provided good results ($R^2 > 0.76$). Spatial analysis was used to draw conclusions from an anthropological / cultural identity point of view. More generally, these techniques not only offer a simple increase in recording capabilities for tangible cultural heritage, they open up new research perspectives, at greater scales.

1. Introduction

The massive introduction of satellite imagery into daily life has modified our perception of space by enabling people to visualize the diversity of our planet, while experiencing perilous exploration from the comfort of their homes. This wealth of documentation, including digital images in visible spectra and beyond, has led to considerable advances in many fields of research [1–4]. This source of knowledge has opened new research perspectives for cultural heritage and archaeology at much vaster scales [5–9]. The digital era has nevertheless produced a bottleneck: human beings may be unable to deal with this huge flow of information in a reasonable amount of time [10–12], particularly when identifying a given structure among a myriad of images. Fortunately, this mundane repetitive task can now be tackled using methods based on machine learning, freeing scientists to devote their expertise to more complex problems [13,14]. Interest in deep learning, a subset of these methods, has increased considerably over the past two decades [15,16], profiting from rapid technical improvements, and key breakthroughs in mathematics, especially in optimization [17]. These techniques are designed to learn from data and make predictions on new instances with a minimum error rate. The architecture of the models is composed of several layers of convolutional artificial neurons, where information flows after being nonlinearly transformed [18,19]. Relevant features are then learnt from a high level of data abstraction [17]. Three main objectives are applied to images: classification [20], object detection [21], and object segmentation [22]. These approaches, based on state-of-the-art deep learning algorithms, have been successfully applyied for

* Corresponding author.
E-mail address: Fabrice.Monna@u-bourgogne.fr (F. Monna).

https://doi.org/10.1016/j.culher.2021.10.004
1296-2074/

proper management and protection of cultural heritage. For example, models have been developed to detect damage to historical buildings automatically [23–24], to map lithology of stones from images [25], or to identify types of weathering in historical stones [26], in order to optimise the choices made in conservation and restoration practices. In this study, we focus on object detection through deep learning, seeking to identify and locate vernacular houses on Sumba Island (Indonesia), from a huge set of satellite images. These traditional houses, known as *rumah adat* in Indonesian, are targeted because they are emblematic of the indigenous local culture [27,28]. They are characterized by a high-pitched central peak in the roof, materializing the connection with the spirits. Ancestral settlements generally contain a few to a few dozen of these houses, organized in circles, or in parallel rows facing each other, together with collective megalithic funeral monuments. Varying in size, these monuments appear to be extensions of clan-houses [29,30]. With economic development and recent cultural globalization, more people are leaving traditional villages, gradually abandoning the traditions, rituals, and religious beliefs of their *marapu* culture [27,30]. These recent mutations disrupt to some extent the traditional way of life and its ancestral cultural foundations.

2. Research aim

Our objective is to propose an easily reproducible workflow allowing human structures to be detected from satellite imagery, and to demonstrate how modern computer vision methods based on deep learning can help to apprehend the on-going complex transformations described above. The implantation of vernacular-style houses, inside and outside traditional villages, is considered an effective proxy. Eight of the most powerful deep learning architectures, belonging to two families of detectors, were compared for their ability to identify and locate traditional houses from a set of satellite images covering the region of Waikabubak, where such houses are abundant. Solutions based on artificial data augmentation were sought to teach the models properly, even with fewer instances. The reason is not only that the labelling phase takes time, but above all because a limited number of examples during the learning phase may present a serious obstacle to the truly effective application of deep learning [12]. Although vernacular houses are plentiful on Sumba Island, such solutions would be very helpful in archaeology, where the available instances may be far scarcer. Once the best approach had been selected, the procedure was extended to the whole of Sumba Island. After separating isolated houses from those belonging to traditional villages, spatial analysis was used to draw conclusions from an anthropological / cultural identity point of view. Several regression algorithms were also applied to estimate the size of the roofs thus identified, as this parameter conveys pertinent information about the economic and political position of the householder.

3. Material and method

3.1. The site

Sumba Island (*Pulau Sumba*), ca. 1500 km ESE of Jakarta, belongs to the Indonesian Lesser Sunda Islands (Fig. 1a). The territory covers ca. 11 000 km^2, with a maximum elevation of 1 225 m asl. Geologically, the island is predominantly composed of sedimentary rocks, with some volcanic / intrusive rocks (Fig. 1b, [31] and references cited therein). The topography of sedimentary formations is mainly coastal terraces and rugged karsts. The tropical dry climate is characterized by seasonal precipitation, abundant only from December to March, with the driest zones along the north and northeast coasts. The population of about 800 000 is mostly rural, concentrated in the hilly fertile western part of the island, except for the largest town of about 40 000 inhabitants [32], Waingapu, located in the northeast (Fig. 1a). The territory is administratively divided into West and East Sumba, corresponding approximately to cultural domains, represented in Fig. 1c by the ethnolinguistic distribution. East Sumba is more linguistically homogeneous than West Sumba. Two large natural parks were created in 1998 on the south coast because of their exceptional biodiversity (Fig. 1a): (*i*) the Manupeu Tanah Daru National Park, covering 870 km^2 and consisting mainly of lowland forests developing on steep slopes, and (*ii*) the Laiwangi Wanggameti National Park, covering 880 km^2 and composed of steppe (60%), and lowland or mountainous rainforest (40%).

3.2. Corpus

Several thousands of typical high-towered houses are present at Sumba, either as isolated buildings or clustered, forming villages and hamlets in the traditional way. Houses are almost square-shaped, from ~6 m to ~20 m in size (Fig. 2). Although originally constructed from pieces of wood and bamboo linked with vegetal ropes, reinforced concrete is often used today. The dense thatch of *alang-alang* grass (a local plant) used to build the roof (Fig. 2a-b) is now increasingly replaced by raw corrugated metal (Fig. 2c; [33]). Note that a few resorts and official buildings imitate the style of these indigenous constructions.

Automatic detection algorithms were evaluated on satellite imagery distributed by Microsoft Bing™ (https://zoom.earth/). The vast majority of images at the highest level of definition available (ca. 0.3 m/px) were neither too dark nor too bright, with almost no clouds. A home-made Python snippet collected tiles 256 × 256 px in size over the targeted geographical extent. They were then merged to produce larger images of 640 × 645 px (corresponding to 187.5 m × 188.5 m), a size suitable for further application of deep learning (Fig. 3a). Reconstructed GeoTIFF images were made to overlap each other by 80 pixels (ca. 25 m) in all directions, to ensure that all houses are complete at least once [34]; see houses 1 and 2 in Fig. 3b, truncated with a blue square, but complete with yellow and red squares. Due to overlap, some complete houses are seen on two or more images (e.g. houses 3 and 4 in Fig. 3b). They were automatically removed using a simple rule: if the centres of their bounding boxes were less than 6 m apart (a value lower than the minimum size of a house), only the most probable item was kept.

Images around Waikabubak (Fig. 1a), within a geographical latitudinal-longitudinal extent of (−9.680°; −9.588°) / (119.360°; 119.470°), were first processed to select the best model. This model was then applied to the whole island with a geographical extent of (−10.346°; −9.271°) / (118.912°; 120.870°). About 700 000 files remained after removing useless offshore tiles.

3.3. Labelling

Before training any object detection model, the operator needs to point out target locations of houses manually in a set of images considered as representative. At that point, two approaches can be applied, using either polygons (and subsequently, models dedicated to segmentation) or simple rectangular bounding boxes (ground-truth boxes), defined by the pixel coordinates of their upper-left and lower-right corners [35]. For the sake of simplicity, we decided to use the latter approach. It is generally admitted that a model is able to detect an object that a human can identify by looking at the image for just 1–2 s. In our case, traditional houses were almost always easy to spot, whatever the roofing material used (Fig. 4), because of their typical square shape, combined with the distinctive shadows produced by the high roof tower. The relatively

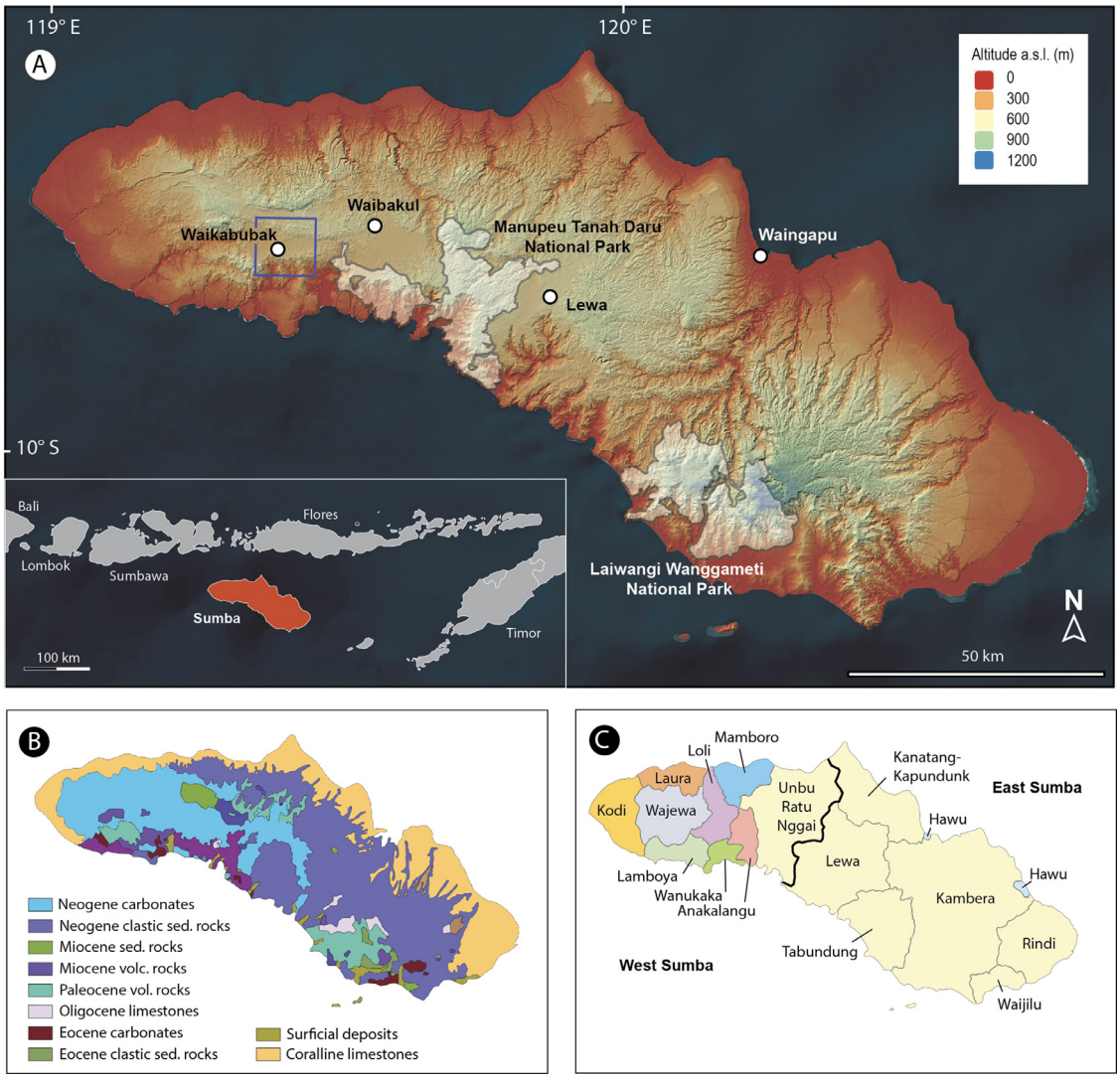

Fig. 1. Geographical, geological, and ethnolinguistic maps of Sumba. (a): hill-shaded, coloured, 30-m resolution digital elevation model (https://www.eorc.jaxa.jp/ALOS/en/aw3d30/); (b): combined bedrock, and superficial geology and age map (http://portal.onegeology.org/OnegeologyGlobal/); (c): approximate geographical locations of the 16 main dialects (Edwards and UBB, 2018).

Fig. 2. Some examples of vernacular houses found at Sumba. (a): collective funeral monuments in the foreground and several vernacular houses in the background; (b) aerial view of a village, with houses exhibiting a high-pitched central peak in their roofs; houses are clustered and organized in rows facing each other because of the limited space on the top of the hill; (c) houses at Mamboro, including a large recent house with a raw corrugated metal roof.

steady downward-facing views from the satellite facilitate this task, although some satellite pictures were also captured in oblique view, and then post-processed by orthorectification before diffusion. This process generated some noticeable deformation, transforming squares into diamonds, but did not affect identification or bounding box positioning (not shown here). When the triangular structure on the roof was not clearly identified, the houses were not labelled (see 1 and 2 in Fig. 4). The annotating step used LabelImg, a free software (https://github.com/tzutalin/labelImg). Note that bounding boxes of houses partly masked by vegetation were kept approximately square by including the supposed position of the building below the canopy. A total of 494 images for the region of Waikabubak produced 1 396 bounding boxes of traditional houses.

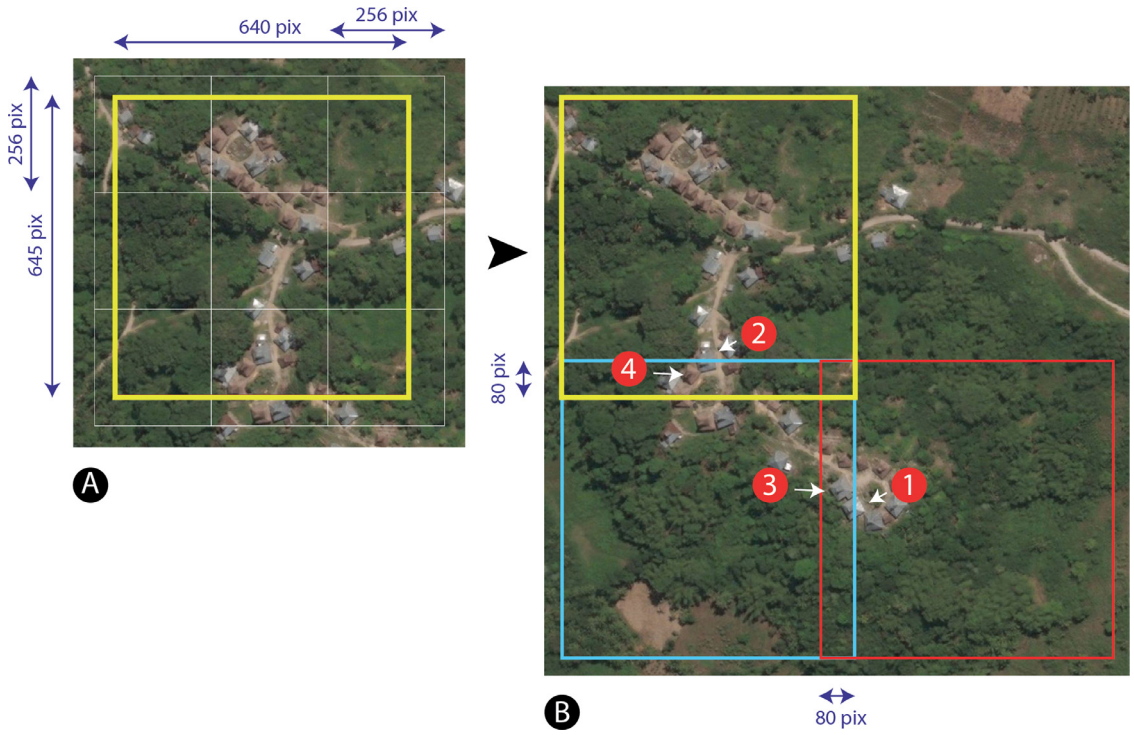

Fig. 3. Composition of images used for further deep learning. (a) images of 640 × 645 px (in yellow) are composed by merging individual 256 × 256 tiles (in white); (b) georeferenced images (in yellow, blue and red) overlap each other by 80 pixels in all directions; objects 1 and 2 are complete in only one image, while objects 3 and 4 are seen completely in two images. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

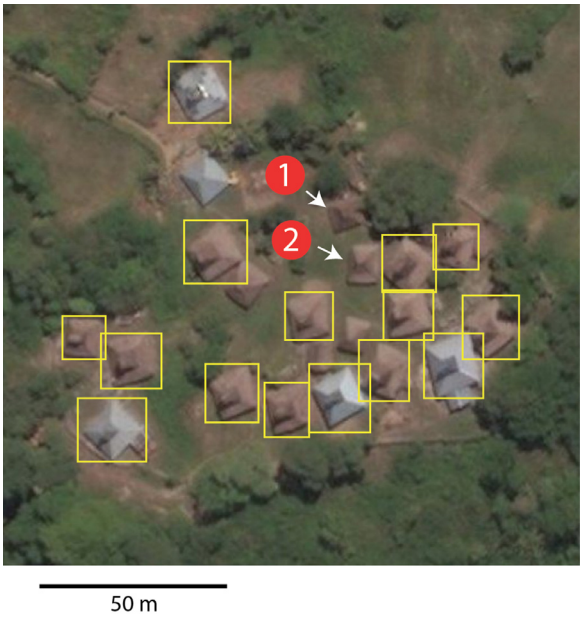

Fig. 4. Final bounding boxes annotated as vernacular houses (in yellow), for one example of a reconstructed image. Manually positioned, they locate the object of focus for deep learning models. The houses with a relatively dark roof summit (see objects 1 and 2), to some extent like those of triangular towers, were not targeted here because they do not present the clear characteristics of true high-towered houses. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.4. Object detection models

Our intention here is not to describe extensively the complex architectures of the models tested, but to provide basic information to explain the underlying principles of deep learning applied to object detection. More details can be found in the abundant specialized literature [17,21,36], many textbooks [37–40], and web sites dedicated to deep learning. Object detectors can be divided into two main categories, one- and two-stage detectors [40,41]. In two-stage detectors, such as Faster R-CNN, the input image is passed through a convolutional neural network (CNN) to obtain a feature map of the image. This part is referred to as the "backbone" network. The map is then used by a Region Proposal Network (RPN), which is a fully convolutional network proposing regions characterized by reference anchor boxes of fixed scales and aspect ratios, placed evenly on the original image. These regions are then filtered by a Non-Maximum Suppression algorithm, whose purpose is to decrease the number of candidate objects to an acceptable level. Bounding box extraction and classification are then obtained for each candidate using regression from the Region of Interest (RoI) pooling layer [42]. The Single Shot Detector (SSD) belongs to the one-stage detector group [43]. It may operate in real-time with a decent trade-off between performance and speed [41,43]. Such a speedy process is obtained by running a convolutional network on the input image only once, and then calculating a feature map. A small convolutional kernel is operated on this feature map to predict bounding boxes and compute classification probabilities. The SSD also uses anchor boxes. It predicts bounding boxes after multiple convolutional layers, which may be of different scales. Results are aggregated, and redundant information is eliminated by applying a non-maximum-suppression algorithm, as with Faster R-CNN. The two-stage detectors may, however, provide better results at the expense of speed. The algorithm schematics of SSD and Faster R-CNN are available in Supplementary SM1, modified after [23] and [44]. Although many of the available models have already been evaluated through various challenges, using for instance the Common Object in Context dataset (i.e. COCO, http://cocodataset.org/#home; [45]), eight different object detection models were tested here: two based on the SSD cat-

Table 1
Performance scores of different architectures, pre-trained on the COCO dataset (models are downloadable at the following address: https://github.com/tensorflow/models/blob/master/research/object_detection/g3doc/detection_model_zoo.md); Training is performed following two configurations including either 480 or 1033 instances of houses. Evaluation is performed on the same 124 instances of houses. No data augmentation is applied. *: stride value between parentheses. See text for definitions of AP@.50:.95, AP @.50, and AP @.75.

Metrics (expressed in%)	AP @.50:.95	AP @.50	AP @.75	AP @.50:.95	AP @.50	AP @.75
Number of instances used for training	480			1033		
Name as in the Tensorflow detection model zoo homepage						
ssd_inception_v2_coco	54.5	95.6	56.6	61.4	98.4	70.0
ssd_resnet_50_fpn_coco	62.0	93.5	76.7	67.2	98.7	84.7
faster_rcnn_inception_v2_coco	53.5	93.6	54.1	59.7	97.1	68.3
faster_rcnn_inception_resnet_v2_atrous_coco (8)*	64.7	96.6	77.2	68.2	98.6	84.9
faster_rcnn_resnet50_coco (16)*	60.1	96.5	70.5	63.6	98.7	78.9
faster_rcnn_resnet50_coco (8)*	65.1	97.1	81.1	68.3	99.5	84.6
faster_rcnn_resnet101_coco (16)*	62.0	95.3	75.7	67.3	97.7	85.2
faster_rcnn_resnet101_coco (8)*	**68.4**	97.4	85.7	**71.9**	98.7	88.7

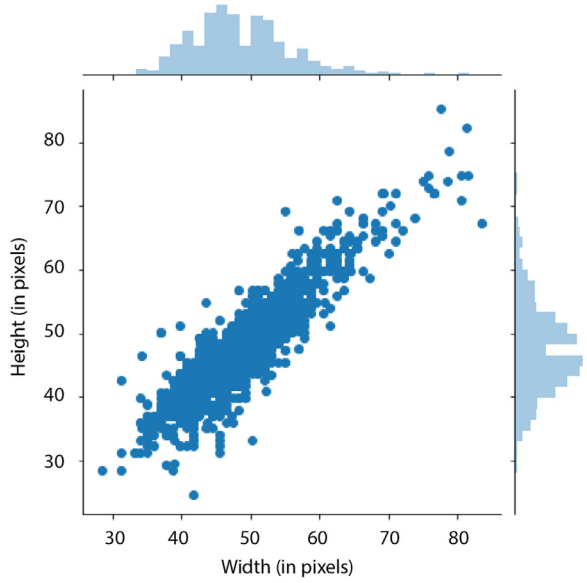

Fig. 5. Shape variation of manually positioned bounding boxes, expressed as height as a function of width (in pixels). The light blue histograms at the top and the right represent the distributions of width and height, respectively.

egory, and six on Faster R-CNN [46]. They are listed in Table 1, with different convolutional neural network backbones, codenamed inception [47], ResNet [48] and inception-ResNet, with or without feature pyramid networks (namely FPN; [49]) as feature extractors. The idea was to reuse these models, already trained for a different task, as the starting point for a custom dataset and specific problem [50]. This approach, known as transfer learning, is expected to speed up the training step and to improve overall performance when sensitive parameters are fine-tuned [38].

3.5. Tuning for house shape and size

For both SSD and Faster R-CNN, a very dense set of potential candidates with different scales and aspect ratios is evenly distributed on the images. The parameters controlling the density, size and shape of these anchor boxes are of primary importance, especially when detecting small objects [51]. An initial guess can be made from the joint distribution of height and width for house bounding boxes observed in the annotated set (Fig. 5). Here, the problem is relatively simple because the bounding boxes will remain square whatever the house orientation (i.e. the houses lie on or around the $y = x$ line in Fig. 5), while their size varies relatively little, between ca. 30^2 and 80^2 pixels (Fig. 5). After applying a k-nearest neighbours' algorithm ($k = 3$), three main sizes can finally be retained {0.15, 0.2, 0.25}, corresponding to 38^2, 51^2 and 64^2 pixels respectively (given a base anchor of 256×256), while two aspect ratios {0.9; 1.1} appeared to be sufficient, since the overall shape is quite regular. Output strides and padding for the extractor are two other important parameters, as they control how deep the abstraction goes to extract features [51]. Denser and more accurate predictions are generally obtained using low strides, but this will notably increase running time [52]. Two values, {8, 16} pixels, were tested here to assess their efficiency in extracting small objects from optical remote sensing images.

3.6. Data augmentation

Deep learning skills are conditioned by the availability of the data feeding the model. While popular datasets used for challenges may contain tens of thousands of images, instances may be much rarer in cultural or archaeological studies. A simple technique, known as data augmentation, has been developed to enlarge the dataset artificially in such cases [53]. This technique supplements the original set of images by new synthetically produced data, obtained by combining geometrical transformations and colour alterations sequentially [54]. Here, six augmentations were tested: 90° rotation, horizontal and vertical flipping, grey level conversion, random colour distortion, and random jitter of box corners by 1 to 4 pixels (see Supplementary Material SM2). A total of 64 experiments (i.e. 2^6 for 6 investigated factors, each with 2 levels: applied / not applied) would be required to evaluate the individual influence of each factor on model performance. As such a full factorial design is time-consuming, factors were selected by building a reduced 2^{6-2} fractional design [55]. Let A, B, C, D be the first four factors; the remaining two: E and F, were chosen so that $E = A^*B^*C$ and $F = B^*C^*D$ (Table 2). This resolution-IV design represents only ¼ of the full 2-level, 6-factor design. Thus, the main effects are not confounded with two-factor interactions, but only aliased with 3-factor and higher-order interactions, which may reasonably be considered insignificant [56].

3.7. Model performance

Three-quarters of the images were randomly selected for training, with the remainder kept for evaluation. Average Precision (AP), a common metric in deep learning, was used to measure the performance of the detectors on the evaluation set. The AP calculation considers the common trade-off between precision and recall, observed at different degrees of correctness for the predicted bounding boxes. These degrees of acceptability are obtained using different thresholds for the Intersection over Union (IoU), a parameter defined as the area of overlap between the predicted box

Table 2
Results of the fractional design (16 carefully chosen experiments) to assess the effect of each individual factor for data augmentation: horizontal and vertical flip of the image, rotation 90°, colour adjustment, grey level conversion, and bounding box jitter. 1 means that the factor is applied (i.e. high level), −1 for not applied (i.e. low level). On the right, the performance scores using as metrics AP@.50:.95, AP@.50, and AP@.75. See text for definition. Experiments were performed using the same 480 and 124 instances of houses for training and evaluation.

	Factor						Metric		
	A Horiz. flip	B Vert. flip	C Rot.	D Colour adj.	E Grey level	F Jitter BB	AP@.50:.95	AP@.50	AP@.75
Exp. #									
1	−1	−1	−1	−1	−1	−1	68.4	97.4	85.7
2	1	−1	−1	−1	1	−1	71.1	98.6	90.7
3	−1	1	−1	−1	1	1	71.7	97.8	89.8
4	1	1	−1	−1	−1	1	70.4	98.5	88.7
5	−1	−1	1	−1	1	1	70.4	98.5	88.6
6	1	−1	1	−1	−1	1	69.4	98.1	87.1
7	−1	1	1	−1	−1	−1	69.9	97.6	85.1
8	1	1	1	−1	1	−1	71.3	98.4	92.0
9	−1	−1	−1	1	−1	1	71.0	97.8	87.2
10	1	−1	−1	1	1	1	70.7	97.3	89.5
11	−1	1	−1	1	1	−1	70.5	97.4	89.3
12	1	1	−1	1	−1	−1	70.8	97.6	88.2
13	−1	−1	1	1	1	−1	70.2	98.2	89.2
14	1	−1	1	1	−1	−1	69.8	98.4	89.1
15	−1	1	1	1	−1	1	70.5	97.6	89.0
16	1	1	1	1	1	1	71.0	97.7	91.5

and the ground truth, divided by the area of their union [57]. The IoU ranges between 0 and 1, with high values indicating more accurate prediction. The metric used is that of the COCO challenge [45], with ten IoU thresholds, from 0.5 to 0.95, at a step of 0.05 (noted AP@.50:.95. AP). It is produced by averaging over the 10 IoU thresholds and tends to reward models that are better at precise localization. In any case, AP@.50:.95 is less generous than IoU $\geq$ 0.5, or the stricter IoU $\geq$ 0.75 limit, which are reported for information.

3.8. Operational settings

For all the experiments, the maximum training epoch was set at 50 000, using an initial learning rate of 3.10^{-4} for the first 10 000 steps, 3.10^{-5} up to 20 000 steps, and then 5 10^{-6}. For Faster R-CNN models, a batch size of 1 (corresponding to a Stochastic Gradient Descent optimization algorithm, [58]) was set, together with a momentum of 0.9, while a batch size of 4 was used for SSD. At a certain point, training loss may continue to decrease, but testing loss may start to increase. This is a clear sign of overfitting, where the model learns well from the training dataset, but fails to generalize this knowledge to make correct inferences on new instances. That is why training was always stopped before 50 000 iterations (often after 15 000 – 45 000). The evolution of the total loss observed on the training set, together with the AP value computed from the evaluation set are reported in Supplementary Information SM3 (using a Faster R-CNN ResNet 101 model).

3.9. Separating houses forming villages from isolated houses

A Density-Based Spatial Clustering of Applications with Noise, DBSCAN [59], algorithm was applied to differentiate isolated houses from those forming villages. When density decreases below a certain point, houses are assumed to lie outside groups. Two sensitive parameters must be set: the minimum number of items necessary for a group, and the maximum distance for clustering with the nearest neighbour. Based on our knowledge of Subanese villages, a 'traditional village' is a cluster of at least three vernacular houses, less than 70 m from their nearest neighbour. Houses that do not fulfil these conditions are defined as 'isolated'. In rare cases, these settings mask the reality of the terrain: in the Loli district, for example, Tarung and Waitabar, two separate villages, were clustered because of their proximity. Without complementary field investigations, such a scenario cannot be identified, but most villages lie hundreds of metres apart.

3.10. Clustering by under- or over-representation of isolated houses

Scan statistics [60], a procedure often used in epidemiology, was applied to examine whether the ratio of isolated houses / houses in villages is the same throughout Sumba territory (i.e. spatial homogeneity) or not (i.e. an underlying geographical structure). Briefly, it consists in scanning the space gradually, centring from one house to another, and counting the number of houses belonging to each type, within expanding circles. Considering a Bernoulli model, a likelihood ratio test is computed for each location and size of the scanning window, using as an alternative hypothesis a high (or low) ratio of isolated (or village) houses within the search window.

3.11. Estimating the surface of recognized houses

With satellite imagery, it is impossible to know the precise house surface area when roofs overhang walls for protection from rainfall. The difference may reach 25%, but roof area is always proportional to house surface area. With houses systematically orientated North – South (or East – West), the problem would be trivial since bounding boxes would approximately match roofs. Here, house orientation is variable. Given its roughly square geometry, a roof can be considered as a square inscribed in another square (bounding box), so that the surface of the roof is comprised between ca. 50% and 100% of the box surface, depending on house orientation. A supervised machine learning approach was developed for further assessment. The distribution of local intensity gradients or edge directions was chosen as input feature. The method, known as Histogram of Orientated Gradients (HOG), has been widely applied for face detection [61]. In brief, it consists in dividing the image into several small cells of $N \times N$ px (Fig. 6a), where gradient intensities and orientations are computed (Fig. 6b). Fig. 6c demonstrates how well local appearance and overall house silhouette are described by gradient distributions, suggesting that this set of variables could be a good candidate to predict roof ori-

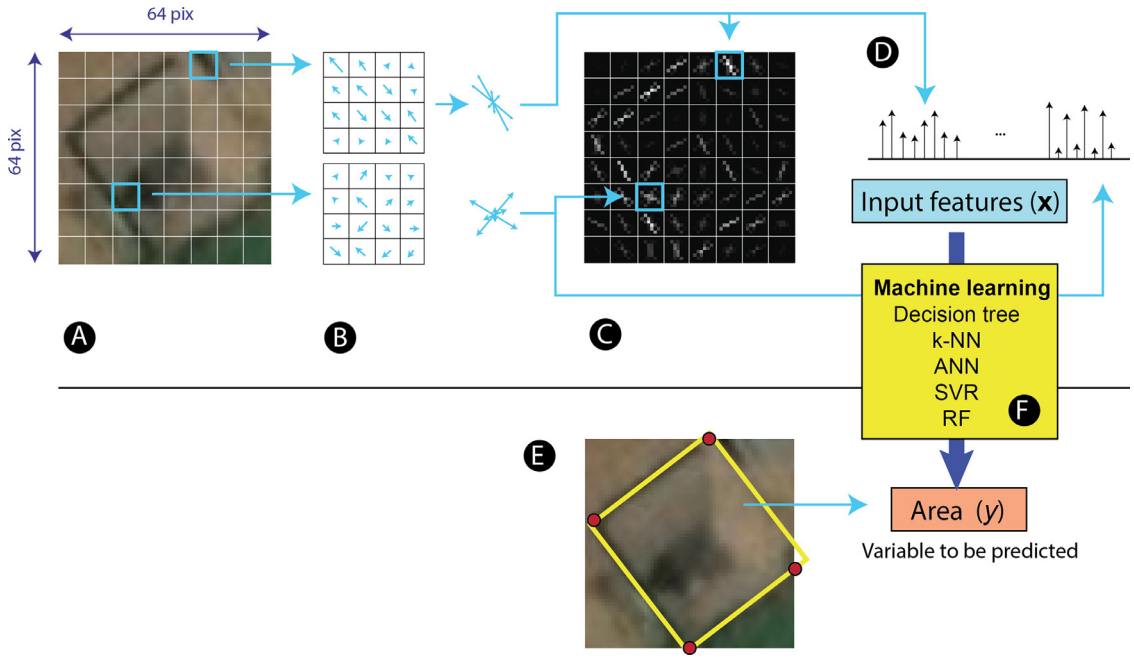

Fig. 6. The machine learning pipeline for assessing the size of vernacular houses, modified after Carcagni, et al. (2015). (a): the image, rescaled at 64 × 64 px, is divided into a 8 × 8 grid; (b): gradients are computed following 12 orientations within every cell; (c): the orientated gradients allow the orientation of the house to be visualized clearly; (d): construction of the input feature vector, **x**; (e): four points manually positioned close to the corners of the roof (in red), and its corresponding minimum bounding rectangles (in yellow), where the surface area is the variable to be predicted: *y*; (f) five machine learning algorithms are applied to input features (a set of oriented gradients). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

entation (see also two additional examples of houses, differently orientated in Supplementary Material SM4). Histograms are then built for each cell of the dense grid covering the image (Fig. 6d). In practice, the images of 363 houses were resized to 64 × 64 px to produce an 8 × 8 grid, where each cell represents 8 × 8 px. A total of 12 orientations was evaluated within each cell, producing an input feature vector, **x**, of 768 values (12 orientations × 64 cells), characteristic of each image (Fig. 6d). In addition, four points were manually placed at (or close to) the four corners of the roof for each of the 363 houses. The surfaces of the minimum bounding rectangles were computed, representing the variable to be predicted: *y* (Fig. 6e). A total of 288 houses was used for training, while the remaining 75 houses were kept for evaluation. Five machine learning algorithms: decision tree, *k*-nearest neighbours, artificial neural networks, support vector machine regression, and random forest, were tested by applying a grid search approach combined with cross-validation to fine-tune the hyperparameters (Fig. 6f; [62]). Their prediction capabilities were evaluated, using coefficients of determination, mean relative errors, and maximum relative errors as quality scores. The best model was then applied to the entire set of houses detected over the Sumba territory.

3.12. Implementation

Object detection models were produced and evaluated using Python 3.7 (https://www.python.org/), and the free TensorFlow object detection API (v. 1.13 including GPU capabilities). Pre-trained models are available at the Tensorflow detection model zoo homepage. The homemade Python snippet for assessing house size relies on the numpy, gdal, scikit-learn, scikit-image, csv, and pandas libraries, and a modified version of the min_bounding_rect.py snippet (https://gist.github.com/kchr/77a0ee945e581df7ed25). Results are expressed as georeferenced polygon vector layers. The homemade snippet for DBSCAN relies on the scikit-learn library [63]. The identification of geographical clusters by scan statistics used the 64-bit SaTScan v9.6 software (http://www.satscan.org/, [60], see user manual for *p*-value calculation with a Monte-Carlo approach).

4. Results and discussion

4.1. Comparing object detection models

The first experiments trained the candidate models using 480 and 1033 instances of traditional houses, and an evaluation set composed of 124 and 363 items, respectively, without any synthetic data augmentation. As expected, performance is better with the largest training set: the gain is ca. 3.5–7% for AP@.50:.95, 1.5–5% for AP@.50, and 3–13% for AP@.75 (Table 1). At first glance, all models provide acceptable outputs, extracting most of the houses, with AP@.50:.95 above 50%, and AP@.50 greater than 90%. Although SSD is fast, it is known to perform less well for small objects, compared with state-of-the-art models based on Faster R-CNN. Using output stride and padding of 8 instead of 16 proves beneficial to the overall achievement of the Faster R-CNN detectors based on ResNet 50 and ResNet 101 (Table 1), as AP@.50:.95 increases by 5–6%, AP@.75 by 4–10%, and AP@.50 by 1–2%. The Faster R-CNN model with a ResNet 101 backbone is retained as it greatly surpasses the other models whatever the number of instances used for training. Despite an increase in running time, an output stride for the extractor of 8 pixels is used, because houses must only be extracted once from satellite images. As no synthetic data augmentation is applied, the performances reported in Table 1 can be considered as the baseline.

4.2. Data augmentation strategy

The question then arises whether the use of an appropriate data augmentation strategy can push significantly further the capabilities for a training dataset of fixed size. The results of the 16 experiments for the fractional design built to assess the effect of each individual factor (i.e., horizontal and vertical flip, rotation, colour

Table 3
Main effect of factors used for data augmentation, i.e. application of random horizontal flip, vertical flip, rotation by 90°, colour adjustment, grey level conversion, and bounding box jitter on AP value. The intercept term corresponds to the last column. The last line corresponds to the best augmentation strategy: 'High' for augmentation applied, 'Low' for not applied.

Factor	A	B	C	D	E	F	Intercept
	Horiz. flip	Vert. flip	rot. 90°	colour adj.	Grey level	Jitter BB	
Main effect (%)	0.13	0.31	−0.14	0.13	0.42	0.20	70.4
Optimal level	High	High	Low	High	High	High	–

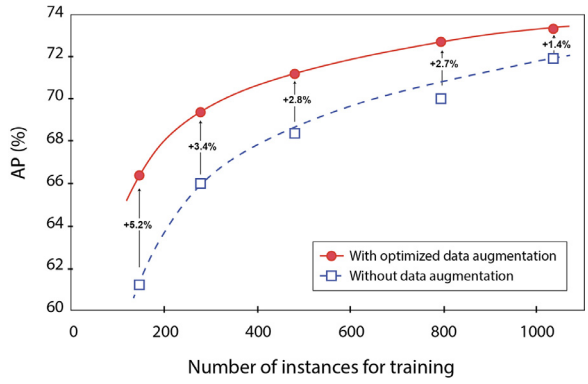

Fig. 7. Evolution of average precision (AP, expressed in%, see text for calculation) as a function of the number of instances used for training. Open blue squares for evaluations made without data augmentation, and red dots for optimal data augmentation; the solid red line and the dashed blue line correspond to interpolated evolutions. The percentages, in black, correspond to the gain obtained using data augmentation for various numbers of instances used for training. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

adjustment, grey level conversion, and jitter of bounding boxes) are presented in Table 2. Note that training involved the lesser of the two configurations used in the previous experiment: the same sets of 480 and 124 instances of houses for training and evaluation, respectively. It appears that AP@.50:.95 increased notably, from no data augmentation (68.4% for exp #1 in Table 2) to data augmentation applying all factors (71.0% for exp #16). Estimates of main effects associated with each factor are reported in Table 3. All factors, except to some extent rotation, therefore seem to be profitable in terms of AP (Table 3). Grey level conversion has the largest (beneficial) effect, followed by vertical flip, and jitter of bounding boxes. The application of horizontal flip and colour adjustment seems to contribute only slightly to AP improvement. It must nevertheless be pointed out that AP is computed with a model frozen after an early stop during the training phase. Identifying this optimal moment is not easy, or at least not perfectly reproducible. Output results (i.e. AP values) may thus suffer somewhat from slight assessment errors of ca. 0.1–0.2%. Given that these uncertainties are of the same order of magnitude as each of the main factor effects (Table 2), it becomes tricky to evaluate with precision their individual contribution to AP. The accumulation of slightly influential factors by applying optimal data augmentation strategy as reported in Table 2 undoubtedly produces a sizable increase in AP values. Such improvements are observed systematically whatever the number of instances used for training (from 144 to 1033; Fig. 7). However, the larger the dataset, the lower the gain, because one cannot afford an explosive increase in terms of AP when numerous instances already illustrate most of the house variability encountered in the field. Note that even when 1033 instances are used for training, the AP value does not reach an asymptotic plateau (Fig. 7). Slight but significant progress might be achieved using more data for learning.

4.3. Inference on all Sumba Island tiles

The final model trained using 1033 instances of houses and data augmentation was used to make inferences on the entire collection of ca. 700 000 tiles covering Sumba Island. After a few days of computation, a total of 22 397 traditional houses was identified, with a confidence score above 0.5 (Supplementary Material SM5). After duplicate removal, 19 143 items remained. Close examination in GIS, using BING satellite imagery as basemap, revealed the presence of several false positives, i.e. objects wrongly identified as traditional houses. These defects were essentially associated with low confidence scores close to (or barely above) 0.5. A cut-off of 0.8 was therefore applied, keeping false negatives (traditional houses missed) at a very low level. A few obvious false positives remained in open spaces, such as rice fields, with isolated trees producing shadows resembling those of house roof towers, or in river talwegs, where shadows from rocks have a similar appearance. After a quick check, most of these mistakes were manually removed, as well as known administrative buildings (e.g. at the airport) and resorts constructed in traditional Sumbanese style. The procedure may nevertheless miss some large houses spotted in the field. These failures are often due to rare poor-quality images, or simply because satellite imagery is not up-to-date (Supplementary Material SM6). Fig. 8a depicts the position of the remaining 14 952 traditional houses throughout Sumbanese territory, after increasing the probability threshold and operating manual cleaning, together with the density gradient (in blue) on the map.

4.4. Identifying traditional villages and isolated vernacular houses

The threshold value of 70 m used to discriminate isolated houses from those belonging to traditional villages is coherent with the abrupt drop observed in the distribution of the distance to the nearest neighbour between houses (see the distribution in Supplementary Material SM7). After calculation, 8 799 houses were considered as isolated (Fig. 8b1 and, at another scale, turquoise dots marked 1 in Fig. 9), while a total of 6 153 houses was assumed to belong to traditional villages (Fig. 8b2, white dots for houses and villages marked 2 in Fig. 9). A total of 1 144 villages, generally with fewer than 5 traditional houses (but up to a maximum of 46), was identified using the above-mentioned rules (see the distribution of the number of houses forming a village in Supplementary Material SM8). Their central positions were then estimated by computing the geographical centroid of village houses (Fig. 8c, yellow stars in Fig. 9). It should be noted that mistakes remain possible: (*i*) the smallest villages will not be identified if they do not contain three traditional houses (see 3 in Fig. 9), (*ii*) two different villages, close to each other, may be grouped together if their nearest houses are less than 70 m apart (see 4 in Fig. 9), (*iii*) several recent houses in the vernacular style, implanted along the main road may be erroneously grouped to form a village, due to their proximity. Such drawbacks, together with the rare mistakes observed during the object detection phase, should not be seen as critical flaws, because the main strength of the

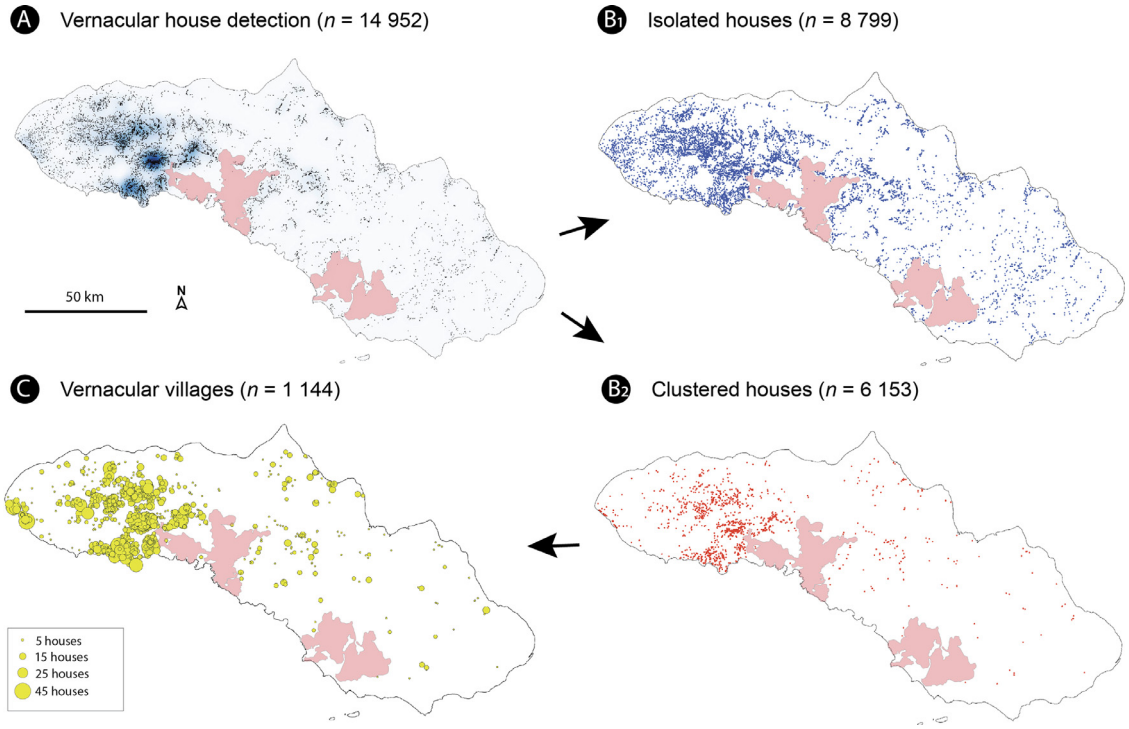

Fig. 8. Vernacular houses and traditional villages identified over the entire Sumba territory. (a): Position of the 14 952 houses identified and density map represented using a blue gradient; darker blue means higher density; (b_1) positions of the 8 799 isolated houses and (b_2) the 6 153 houses forming traditional villages; (c) position of the 1 144 traditional villages, with symbol size varying in relation to the number of vernacular houses forming the village. The pale pink polygons correspond to the two national parks.

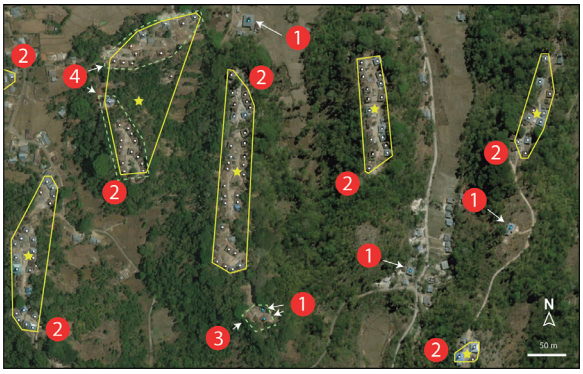

Fig. 9. Close-up in the region of Waikabubak depicting the identified vernacular houses, either isolated (turquoise dots) or organized in traditional villages (white dots); 1 for isolated houses; 2 for villages (yellow polygons, yellow star for centroid); 3 for a traditional village missed because it only contains two vernacular houses; 4 for two separate villages erroneously combined into one single entity. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

approach is that it tends toward exhaustivity over a large territory; these potential shortcomings will have no effect on statistical analysis.

Scan statistics demonstrate spatial structuration: the ratio of isolated (or village) houses is clearly not homogeneous throughout the study area. Three main clusters, which cannot be attributed to random effects, are statistically recognized (Fig. 10a). The first, centred on Waikabubak, presents a higher ratio of houses in villages (in blue in Fig. 10a). The second covers East Sumba almost entirely, with a higher-than-expected ratio of isolated houses (in red in Fig. 10a). The third, also with a higher ratio of isolated houses, is located at the extreme west of the island (in red in Fig. 10a).

Table 4
Algorithm performance scores: decision tree, *k*-NN for *k*-nearest neighbour regression, neural network, SVR for support vector machine for regression, and random forest; R^2 for coefficient of determination of the prediction, MRE for mean relative error, and MaxRE for maximum relative error. Several sets of feature variables were tested: the full set, namely without PCA, and four reduced sets allowing 50%, 65%, 80%, and 95% of the total variance to be explained. In bold, the best result obtained, corresponding to the method used.

	Without PCA	With PCA			
Variance explained (%)	100	95	80	65	50
Decision tree					
R^2	0.704	0.775	0.775	0.761	0.764
MRE (%)	8.4	7.3	7.3	7.0	7.5
MaxRE (%)	26.9	20.9	20.9	24.5	20.0
k-NN					
R^2	0.859	0.855	0.845	0.845	0.859
MRE (%)	5.7	5.6	5.7	5.8	5.5
MaxRE (%)	17.7	18.0	19.4	19.3	19.6
Neural network					
R^2	0.673	0.734	0.813	0.858	0.840
MRE (%)	8.7	6.6	5.9	5.6	5.9
MaxRE (%)	21.8	25.9	15.9	15.3	18.7
SVR					
R^2	0.787	0.848	0.868	**0.883**	0.870
MRE (%)	6.6	5.7	5.5	**5.3**	5.1
MaxRE (%)	17.7	14.7	14.9	**13.6**	16.5
Random Forest					
R^2	0.793	0.839	0.837	0.812	0.818
MRE (%)	6.7	6.1	6.1	6.4	5.9
MaxRE (%)	19.3	16.9	17.3	20.1	18.3

4.5. House size

Table 4 presents the results of the five machine learning algorithms tested, and their performance in assessing house size from oriented gradient values. This was computed using either the complete raw dataset as input, or a reduced set obtained by Princi-

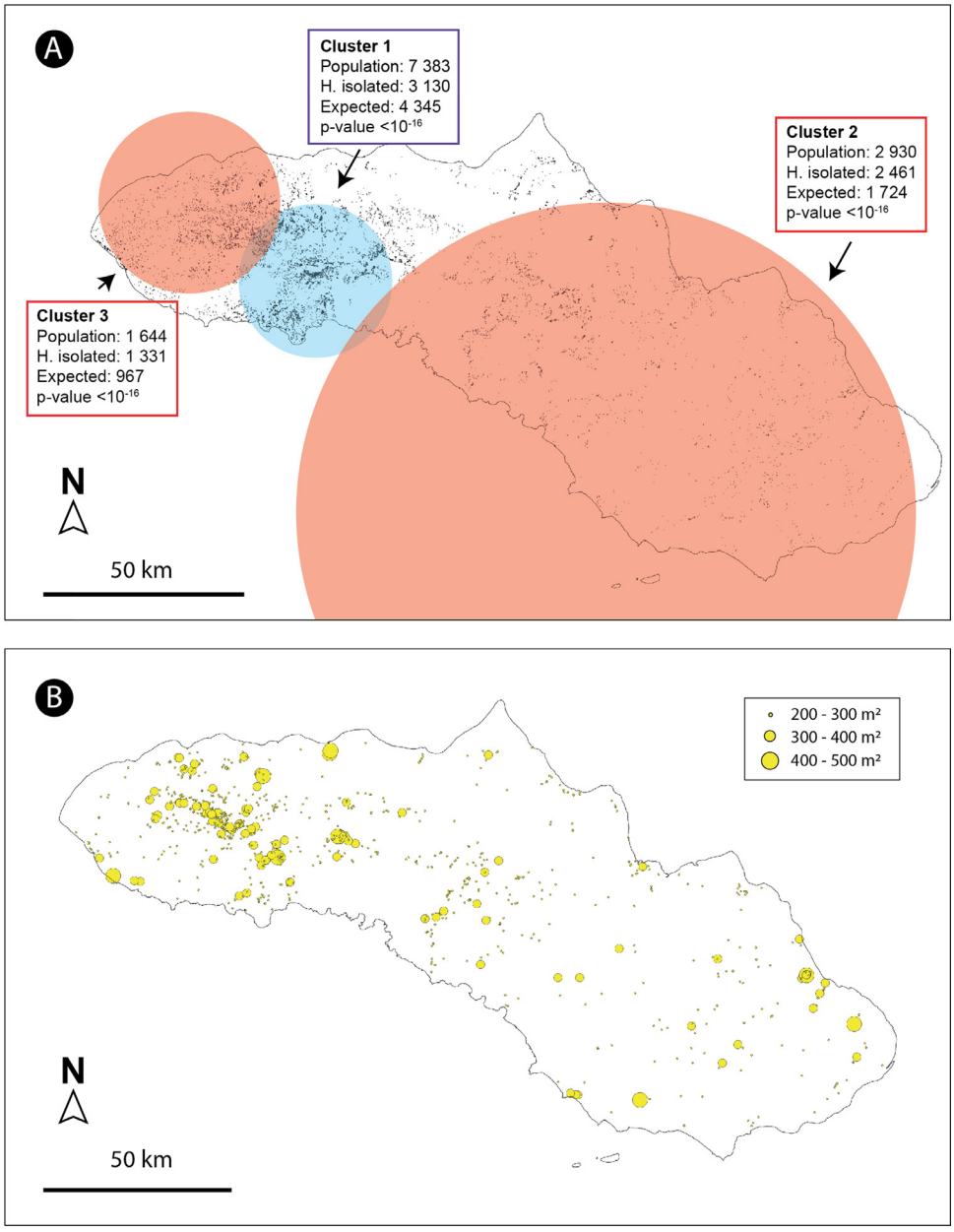

Fig. 10. (a): Results of the scan statistics processing over the entire Sumba territory, for all identified isolated houses, and those forming villages. The blue circle represents a cluster with a high ratio of vernacular houses belonging to villages; the two red circles correspond to areas with high ratios of isolated houses. For each cluster, the total number of houses within the circles, the number of isolated houses, the expected number under the null hypothesis (spatial homogeneity), and the *p*-value of the existence of a cluster, are provided. (b): Spatial distribution of house roofs larger than 200 m^2. The size of the circles varies with roof size. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

pal Component Analysis (PCA) to capture maximum total variance with lower dimensionality. Using the entire set of 768 values as input data (i.e. without PCA transformation), performance scores are satisfactory, with high R^2 values, varying from 0.67 to 0.86, whatever the method. The best results are obtained with *k*-nearest neighbours. The score of all models tends to improve with data reduction, except for the *k*-nearest neighbour algorithm that remains steady. The biggest boost is observed for the artificial neural network, with a notable R^2 increase of almost 0.2. This score is slightly lower than that of the support vector machine algorithm with PCA retaining 65% of the explained variance, where R^2 reaches almost 0.9, while the mean and maximum relative errors are around only 5% and 14%, respectively (see Supplementary Material SM9 for comparison between ground truth and predicted surface). Note, however, that approaches based on mask R-CNN, which extract pixels belonging to the objects of interest, could also be used to determine house size. This option was not chosen here, but the results presented above indicate that oriented gradients are also quite efficient to estimate house size rapidly and accurately, given that errors of only ca. 15% (at the most) are perfectly acceptable in our case. Roof size varies between 50 and 440 m^2, with a mean of 136 m^2 (see Supplementary Material SM10). A total of 1 027 roofs exceeded 200 m^2 in surface area (Fig. 10b). Interestingly, the size distribution of isolated houses appears to be statistically different from those belonging to villages, with larger roofs for isolated houses (KS = 0.1, *p*-value $< 10^{-16}$ for 2-sample Kolmogorov-Smirnov statistic, see also Supplementary Material SM11 depicting cumulative distributions for both groups).

4.6. Cultural implications

Most traditional houses are in West Sumba (~ 80%), especially in the Loli and Wanukaka territories, and to some extent in the Wajewa area (see Figs. 8a and 1c for ethnolinguistic subdivisions). This distribution tends to follow the implantation of the Sumbanese population, whether the inhabitants live in traditional houses or not (except for the large town of Waingapu). The greatest house density is found over Neogene carbonates (Fig. 1b), in particular on the hilly terrains surrounding the rice fields of the Waikabubak and Waibakul lowlands, where fertile soils must have attracted populations. In the Loli and Wanukaka territories, ancestral clan-house settlements were often established at the top of hills for defensive purposes ([64]; see Figs. 8c and 9 depicting an area located to the south of the Waikabubak lowland, where villages are massively implanted along the ridges). Another good reason for building there is that the air is drier than in the lowlands, close to the rice fields [29]. Some large villages are also found in the Kodi territory, located to the extreme west of the island (Fig. 8c), where the inhabitants exploit the few agricultural plots developed on coralline limestones (Fig. 1b). Other small settlements are present in the western part of the island, but they have not been identified here, because they are essentially composed of normal (not high-pitched) houses, which do not fulfil the criteria to label them as traditional. Today, there is no need for defence, so that isolated houses are quite numerous. In West Sumba, they mostly correspond to buildings in medium-sized villages, or along the main communication routes (Fig. 8b$_1$). With the rising standard of living, people have progressively left ancestral settlements for modern accommodation and related services (e.g. regular water supply, salaried employment, etc.). Others have built their homes closer to the fields they exploit for practical reasons. These modern houses ("garden houses") are generally larger and more comfortable than those originally found in traditional villages. However, the break with tradition is only partial as these new houses are frequently built close to the original village of their owner, and often imitate traditional Sumbanese style. Around Waikabubak, the ratio of isolated houses is lower than over the rest of the island (Fig. 10a), suggesting that the ancestral structure of society is more prevalent there. In East Sumba, the situation is very different. Traditional houses are scarcer (Fig. 8a), except to some extent in the region of Lewa, where numerous agricultural fields are exploited. The dominance of Neogene clastic sedimentary rocks as bedrocks (Fig. 1b), combined with steep slopes (Fig. 1a) and low rainfall makes the terrains of East Sumba not very productive and difficult to exploit. Even where traditional houses are present, they are often scattered in response to environmental constraints, so that only a few traditional villages can be identified (Fig. 8c). The political functions of these villages were, however, not the same as in the western part of the island. In the early 20th century, travellers reported that a single aristocrat assisted by a few vassals reigned over the eastern part of the island, while no such centralized power existed in the western domain. The sovereign ruled over a strongly stratified society, where possessing slaves was a sign of a high-ranking social and economic position. Note that slavery was abolished in 1860 in the Dutch East Indies, but castes still persist today, despite the efforts of the democratic Indonesian government. The villages and their activity are still centred on these aristocratic houses, whose size and position are clearly related to the social standing of their owners, their prestige, political power, wealth, and connection to the spirits (see as an example Fig. 2c, depicting an aristocratic house in the foreground of the picture). There are virtually no vernacular houses within the two large natural parks established in 1998 on the south coast (Fig. 8), which are predominantly established on volcanic substrates (Fig. 1b). Yet some ancient collective graves, testifying to the presence of former villages, have been identified here and there. This area has a low agricultural potential, and the creation of the two protected areas has probably incited the remaining population to abandon the area and to move elsewhere.

5. Conclusion

Applying an appropriate set of techniques based on machine learning can help to bridge the gap between tangible and intangible heritage, by rapidly producing maps for further spatial analyses undertaken with specific objectives. Nevertheless, it should be mentioned that the overall quality of satellite imagery is a strong limiting factor for applying deep learning in good conditions. Here, images were neither too dark nor too bright, with almost no cloud, always remaining perfectly readable at the highest resolution available (about 0.3 m/px), which is quite sufficient for our purpose. Some of the examples provided, however, demonstrate that the period during which satellite images have been acquired may introduce errors, because the situation is continuously evolving. Note also that if the labelling phase is inadequate or not fully representative, the accuracy of detection will be greatly impaired. Whatever the precautions taken, the techniques described here inevitably introduce some false positives and false negatives, which the researcher will have to manage. Automatic clustering of houses to identify villages implies a choice of settings. Even if the settings are appropriate, errors may also be introduced at that step. That is why a solid knowledge of the context is always highly desirable. Returning to the field for verification or for further investigation is an option that should never be neglected.

Our attempts to explain the distribution of vernacular houses are rendered complex as both environmental and societal factors almost certainly play a role, in competition with recent developments in Sumbanese society that are reshuffling the cards. The growing importance of 'garden houses' merits particular attention. It could be interesting to examine whether the situation of isolated houses without towers is similar, as differences in architecture may reflect the personal relationship of the owner to tradition. Further research could complete the data with surveys over time, or could repeat the study on available aerial images, prior to the societal changes mentioned.

Technically, several types of object detectors are available but, beyond this crucial choice, the researcher must be aware that other important parameters may heavily impact the overall performance of the model (output stride, augmentation strategy, etc..). The number (and diversity) of instances used for training is among these sensitive variables. In the present case, the experiment shows that using more than 1000 instances probably does not reach the maximum performance score. Nonetheless, it would have been absurd to increase the size of the training set to the point of labelling almost all the houses that must be detected. This issue is a matter of balance, which must be adapted to the problem to be solved. Interestingly, using an appropriate deep learning model for object detection, followed by a machine learning approach with HOG values as feature inputs, the size of houses can be estimated with good accuracy. Although the proposed workflow is sufficiently efficient here, instance segmentation, which straightforwardly extracts pixels belonging to houses, could also be evaluated in the future.

In the light of this summary, a novice might be afraid of the technicity required and the potential difficulties related to practical implementations. It would however be an error because all tools, freely available, come with detailed documentation. Even though time will be required during development, the gain will be huge for objects to be detected at very large scales, an impossible task if done by hand. In the present example, only one single class was involved, but it is entirely possible to detect several objects, at one

and the same time. The potential of deep learning applied to various situations related to the recording (and thus to the preservation or the study of the dynamics) of tangible cultural heritage is huge. These techniques not only offer a simple increase in recording capabilities, they provide a whole new order of magnitude for research perspectives.

Acknowledgements

This research was funded by the Conseil Régional de Bourgogne – Franche Comté. We are grateful for comments by the anonymous reviewers and by the editor, which have greatly improved the manuscript.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.culher.2021.10.004.

References

[1] T.J. Schmugge, W.P. Kustas, J.C. Ritchie, T.J. Jackson, A. Rango, Remote sensing in hydrology, Adv. Water Resour. 25 (2002) 1367–1385.
[2] D.S. Boyd, F.M. Danson, Satellite remote sensing of forest resources: three decades of research development, Prog. Phys. Geogr. 29 (2005) 1–26.
[3] N. Horning, J.A. Robinson, E.J. Sterling, W. Turner, S. Spector, in: Remote Sensing for Ecology and Conservation: a Handbook of Techniques, OUP Oxford, 2010, p. 496.
[4] R.P. Gupta, in: Remote Sensing Geology, 3rd ed., Springer, 2017, p. 428.
[5] S.H. Parcak, in: Satellite Remote Sensing For Archaeology, Routledge, 2009, p. 320.
[6] G. Caspari, Mapping and damage assessment of "Royal" burial mounds in the Siberian Valley of the Kings, Remote Sens. 12 (2020) 773.
[7] L. Luo, X. Wang, H. Guo, R. Lasaponara, X. Zong, N. Masini, G. Wang, P. Shi, H. Khatteli, F. Chen, S. Tariq, J. Shao, N. Bachagha, R. Yang, Y. Yao, Airborne and spaceborne remote sensing for archaeological and cultural heritage applications: a review of the century (1907–2017), Remote Sens. Environ. 232 (2019) 111280.
[8] G. Caspari, P. Crespo, Convolutional neural networks for archaeological site detection – Finding "princely" tombs, J. Archaeol. Sci. 110 (2019) 104998.
[9] D.S. Davis, Geographic disparity in machine intelligence approaches for archaeological remote sensing research, Remote Sens. 12 (2020) 921.
[10] A. Traviglia, A. Torsello, Landscape pattern detection in archaeological remote sensing, Geosciences (Basel) 7 (2017) 128.
[11] F. Monna, J. Magail, T. Rolland, N. Navarro, J. Wilczek, J.O. Gantulga, Y. Esin, L. Granjon, A.-.C. Allard, C. Chateau-Smith, Machine learning for rapid mapping of archaeological structures made of dry stones – example of burial monuments from the Khirgisuur culture, Mongolia, J. Cult. Herit. 43 (2020) 118–128.
[12] F. Emmert-Streib, Z. Yang, H. Feng, S. Tripathi, M. Dehmer, An Introductory review of deep learning for prediction models with big data, Front. Artif. Intell. 4 (2020) 1–23.
[13] K. Lambers, W.B. Verschoof-van der Vaart, Q.P.J. Bourgeois, Integrating remote sensing, machine learning, and citizen science in Dutch archaeological prospection, Remote Sens. 11 (2019) 794.
[14] M. Soroush, A. Mehrtash, E. Khazraee, J.A. Ur, Deep learning in archaeological remote sensing: automated Qanat detection in the Kurdistan region of Iraq, Remote Sens. 12 (2020) 500.
[15] G.E. Hinton, S. Osindero, Y.-.W. Teh, A fast learning algorithm for deep belief nets, Neural Comput 18 (2006) 1527–1554.
[16] A. Krizhevsky, I. Sutskever, G. Hinton, ImageNet classification with deep convolutional neural networks, Proc. Adv. Neural Inf. Process. Syst. 25 (2012) 1090–1098.
[17] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444.
[18] S. Lawrence, C.L. Giles, A.C. Tsoi, A.D. Back, Face recognition: a convolutional neural-network approach, IEEE Trans. Neural Netw. 8 (1997) 98–113.
[19] V. Nair, G.E. Hinton, Rectified linear units improve restricted Boltzmann machines, in: Proceedings of the 27th International Conference on Machine Learning (ICML-10), 2010, pp. 807–814.
[20] T. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, Y. Ma, PCANet: a simple deep learning baseline for image classification? IEEE Trans. Image Processing 24 (2015) 5017–5032.
[21] Y. Xiao, Z. Tian, J. Yu, J.X. Wu, A review of object detection based on deep learning, Multimed. Tools Appl. 79 (2020) 23729–23791.
[22] K. He, G. Gkioxari, P. Dollar, R. Girshick, Mask r-cnn, 2017, arXiv:1703.06870.
[23] N. Wang, X. Zhao, P. Zhao, Y. Zhang, Z. Zou, J. Ou, Automatic damage detection of historic masonry buildings based on mobile deep learning, Autom. Construct. 103 (2019) 53–66.
[24] Z. Zou, X. Zhao, P. Zhao, F. Qi, N. Wang, CNN-based statistics and location estimation of missing components in routine inspection of historic buildings, J. Cult. Herit. 38 (2019) 221–230.
[25] M.E. Hatır, İ. İnce, Lithology mapping of stone heritage via state-of-the-art computer vision, J. Build. Eng 34 (2021) 101921.
[26] M.E. Hatır, M. Barstuğan, İ. İnce, Deep learning-based weathering type recognition in historical stone monuments, J. Cult. Herit. 45 (2020) 193–203.
[27] C. Jeunesse, Sacrifice et partage dans l'île de Sumba (Indonésie), L'archéologue 150 (2019) 66–69.
[28] M. Devanastya, The transformation of form and discourse of identity in Sumbanese houses and settlements, in: Proceedings of the 3rd International Conference on Dwelling Form (IDWELL 2020), 2020, pp. 149–160.
[29] R.L. Adams, A. Kusumawati, The social life of tombs in West Sumba, Indonesia. In Adams, K. L., and King, S. M. (eds.), Residential Burial: A Multiregional Exploration, Archeological Papers No. 20, American Anthropological Association, Wiley, Hoboken, NJ, (2010) 17–32.
[30] C. Jeunesse, Dualist socio-political systems in South East Asia and the interpretation of late prehistoric European societies, in: Habitus? The Social Dimension of Technology and Transformation, Edited by Sławomir Kadrow & Johannes Müller, Sidestone Press Academics, Leiden, 2019, 181–213.
[31] C.I. Abdullah, J.-.P. Rampnoux, H. Bellon, R.C. Maury, R. Soeria-Atmadja, The evolution of Sumba Island (Indonesia) revisited in the light of new data on the geochronology and geochemistry of the magmatic rocks, J. Asian Earth Sci. 18 (2000) 533–546.
[32] Badan Pusat Statistik Kabupaten Sumba Timur, Kota Waingapu Dalam Angka, 2020, 128 pp.
[33] J.W. Mross, Cultural and architectural transitions of Southwestern Sumba island, Indonesia. Acs4 2000 International Conference, 260–265.
[34] Y. Koga, H. Miyazaki, R. Shibasaki, A method for vehicle detection in high-resolution satellite images that uses a region-based object detector and unsupervised domain adaptation, Remote Sens. 12 (2020) 575.
[35] A. Géron, in: Hands-On Machine Learning With Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques, 2nd Ed., O'Reilly, 2019, p. 566.
[36] K.L. Masita, A.N. Hasan, T. Shongwe, Deep learning in object detection: a review, in: 2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD), Durban, South Africa, 2020, pp. 1–11.
[37] I. Goodfellow, Y. Bengio, A. Courville, in: Deep Learning, MIT Press, 2016, p. 800.
[38] A. Gulli, A. Kapoor, S. Pal, in: Deep Learning with TensorFlow 2 and Keras: Regression, ConvNets, GANs, RNNs, NLP, and More With TensorFlow 2 and the Keras API, 2nd Edition, Packt Publishing, 2019, p. 646.
[39] A. Géron, in: Deep Learning avec Keras et TensorFlow - 2e éd. - Mise en œuvre et Cas concrets: Mise en œuvre et Cas Concrets, 2nd Ed., O'Reilly, 2020, p. 576.
[40] S. Agarwal, J.O. Du Terrail, F. Jurie, Recent advances in object detection in the age of deep convolutional neural networks, ArXiv:1809.03193.
[41] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, M. Pietikäinen, Deep learning for generic object detection: a survey, Int. J. Comput. Vis. 128 (2020) 261–318.
[42] J. Dai, Y. Li, K. He, J. Sun, RFCN: object detection via region based fully convolutional networks, in: NIPS, 2016, pp. 379–387.
[43] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, SSD: single shot multibox detector, 2015, arXiv:1512.02325v2.
[44] A.N. Veeranampalayam Sivakumar, J. Li, S. Scott, E.J. Psota, A. Jhala, J.D. Luck, Y. Shi, Comparison of object detection and patch-based classification deep learning models on mid- to late-season weed detection in UAV imagery, Remote Sens. 12 (2020) 2136.
[45] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, L. Zitnick, Microsoft COCO: common objects in context, in: ECCV, 2014, pp. 740–755.
[46] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, in: NIPS, 2015, pp. 91–99.
[47] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–15.
[48] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[49] T.-.Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.
[50] K. Weiss, T.M. Khoshgoftaar, D. Wang, A survey of transfer learning, J. Big Data 3 (2016) 9.
[51] Y. Ren, C. Zhu, S. Xiao, Small object detection in optical remote sensing images via modified Faster R-CNN, Appl. Sci. 8 (2018) 813.
[52] J. Yan, H. Wang, M. Yan, W. Diao, X. Sun, H. Li, IoU-adaptive deformable R-CNN: make full use of IoU for multi-class object detection in remote sensing imagery, Remote Sens. 11 (2019) 286.
[53] D.M. Montserrat, Q. Lin, J. Allebach, E.J. Delp, Training object detection and recognition CNN models using data augmentation, in: IS&T International Symposium on Electronic Imaging, 2017, pp. 27–36.
[54] C. Shorten, T.M. Khoshgoftaar, A survey on image data augmentation for deep learning, J. Big Data 6 (2019) 60.
[55] J.L. Goupy, in: Methods for Experimental Design, Elsevier, 1993, p. 465.
[56] T. Lundstedt, E. Seifert, L. Abramo, B. Thelin, A. Nyström, J. Pettersen, R. Bergman, Experimental design and optimization, Chemometr. Intell. Lab. Syst. 42 (1998) 3–40.
[57] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, S. Savarese, Generalized intersection over union: a metric and a loss for bounding box regression, CVPR, 2019.

[58] L. Bottou, Large-Scale machine learning with stochastic gradient descent, in: Proc. COMPSTAT'2010, 345, 2010, pp. 177–186.
[59] M. Ester, H.-.P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, 1996, pp. 226–231.
[60] M. Kulldorff, Spatial scan statistics: models, calculations, and applications. In Recent Advances on Scan Statistics (eds J. Glaz and N. Balakrishnan), Boston: Birkhauser. (1999) 303–322.
[61] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: International Conference on Computer Vision & Pattern Recognition (CVPR '05), San Diego, United States, Jun 2005, pp. 886–893.
[62] S. Raschka, V. Mirijalili, in: Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-Learn, and TensorFlow, 2nd Ed., Packt Publishing, 2017, p. 622.
[63] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[64] A. Denaire, C. Jeunesse, F. Monna, L. Waldvogel, Quelques remarques sur les enceintes en pierre sèche des habitats traditionnels actuels de l'île de Sumba (Indonésie), in: le phénomène des enceintes dans le Néolithique du nord-ouest de l'Europe, 33e colloque interrégional sur le Néolithique, Saint-Dié-des-Vosges, 8-9 novembre 2019, in press.