Real-Time Identification of Artifacts: Synthetic Data for AI Model

Andrea Tomalini
Edoardo Pristeri

Abstract

The collections represent the constitutive element and the *raison d'être* of each museum. Their management, care and dissemination are therefore a task of primary importance for every museum. Applying new Artificial Intelligence technologies in this area could lead to new initiatives. However, the development of certain tools requires structured and labeled datasets for the training phases which are not always easily available. The proposed contribution is within the domain of the construction of specific datasets with low budget tools and explores the results of a first step in this direction by testing algorithms for the recognition and labeling of heritage objects. The developed workflow is part of a first prototype that could be used both in heritage dissemination or gamification applications, and for use in heritage research tools.

Keywords
heritage, museum collections, image recognition, synthetic dataset, machine learning.

doi.org/10.3280/oa-845-c216

Introduction

Museum collections represent a part of our cultural heritage and as such are testimony of our past, enrich the present and inform the future. This is sufficient motivation to establish that it is essential to increase the perceived value of these artifacts also through new methodologies. Using Artificial Intelligence (AI) techniques, is however a rather challenging task: there are relatively few datasets to use for smart applications, and the metadata associated with images is often poor and unstructured, or does not exist.

To answer the last of these issues, this paper presents a first prototype which, starting from photogrammetric acquisitions, builds synthetic datasets of correctly labeled images to train AI algorithms capable of labeling museum objects.

AI frameworks capable of classifying image contents have already achieved some success in various fields, these can be conceptualized as memories capable of being recalled to perform a certain task. It follows that the construction of a memory is the most delicate phase for the development of these algorithms and it is essential, whether we are talking about Machine Learning or Deep Learning, to have access to a large data set of labeled images on which to perform the training phases. As for everyday objects there are numerous sets available, the same cannot be said for heritage artifacts. In this case it is necessary to acquire the data and then label it, these operations are quite long and, consequently, expensive. The proposed prototype – born from the collaboration of a research group of the Department of Architecture and Design of the Politecnico di Torino, a working group of the LINKS Foundation and the Museo Egizio of Turin – suggests an economic method for the creation of a labeled image dataset automatically. As anticipated, the proposed workflow is based on the acquisition of the artifacts through photogrammetric survey techniques. The high-fidelity textured mesh of the object obtained is processed by Physically Based Rendering tools for the creation of image datasets for training the algorithms.

Related Works

The development of high-performance computing systems, advances in the design of the software architecture structure and, above all, the availability of large photographic datasets, have been the main factors that have pushed the classification and detection methods to success. On AI systems, datasets such as MS COCO with more than 300,000 images [Lin et al. 2014], have certainly contributed to the research and development of these frameworks. However, photographic datasets are not always easily available, especially if they are datasets concerning objects not in common use.

To date, with regard to semantic segmentation in the context of cultural heritage, there is still a scarcity of significant photographic datasets. Examples have been reported in the bibliography in which, to overcome this problem, the transfer learning technique was applied. This technique consists in using large and generic datasets (such as the one previously cited) along with a reduced set of images containing the object target of the recognition problem. [Marinescu et al. 2020].

The use of 3D models for the construction of synthetic datasets for image classification is a fairly recent area of research and appears to be a rather promising technique, complementary to the one described above [Nowruzi et al. 2019] [1]. In the bibliography there are some examples in which researchers have used synthetic datasets to train neural networks, for example NVIDIA researchers have created 60,000 annotated photorealistic renders of objects that are in free fall now collected within the open "Falling Things" (FAT) [Tremblay et al. 2018]; several researches were conducted to test the goodness of synthetic datasets for the recognition of objects in a domestic or commercial environment, in this case the physical environment surrounding the object to be recognized was also analyzed and (therefore taking into account occlusions or positions random and unusual object and observer's chamber), in all cases, it was underlined how the results obtained are promising and how it is much more important to generate random images that simulate more the randomness of the physical environment, rather than photorealistic [Mitash et al. 2017; Rajpura et al. 2017; Hinterstoisser et al. 2019]. The semantic segmentation problem [2] has also started to benefit from these synthetic datasets, with the creation of virtual scenes to perform the segmentation of indoor environments [Handa et al. 2015; Papon et al. 2015]. On the basis of this research and strengthened by the previous research experience [Tomalini et al. 2021] the group has re-designed the methodology of construction of the datasets to reduce the problem of overfitting caused by the reduced variability of the developed synthetic dataset [3] and to increase the overall robustness.

Case Study

In collaboration with the researchers of the Museo Egizio of Turin, two pairs of vases were chosen as a case study: the first pair is labeled as Class B, Black Topper Pottery, belonging to the Predynastic Period, Naquada I (4500-3100 BC); the second pair is labeled as Class R Rough Faced of the Old Kingdom, III-IV dynasty (2680-2140 BC).

Given the initial state of the research, these artifacts were selected because the classes to which they belong are identifiable by peculiarities that are rather easy to recognize. It follows that the classification algorithms used for these artifacts in this work are simple and based on a reduced number of characteristics. The Class B, Black Pottery is characterized by this crown of black pigment at the mouth and lip of the vessel, the rest of the body is red. Class R, Rough Faced is characterized by the absence of decorations and by its very rough ocher-colored surface.

Data Acquisition

As anticipated, the vessels were detected through photogrammetric techniques. However, to assign a correct scale to the model, and consequently ensure a good rendering, the cloud was scaled on the basis of known points (belonging to the artifact such as: cracks, imperfections or stains) previously acquired with the aid of a scanner. structured light. The instrument used, with accuracy from the technical data sheet of 0.3mm, is not able to return a faithful texture and for this reason we have preferred to use an integrated approach. The instrumentation used was a Sony Alpha 6000 equipped with a 23.5x15.6 mm sensor and a first generation Revo Point POP structured light scanner. Despite the unprofessional and low cost tools, it is known in the scientific community that the processes of generating point clouds from georeferenced photogrammetric blocks provide excellent results even when the starting data is not a set of images acquired with a calibrated photogrammetric camera [Cardenal et al. 2004].

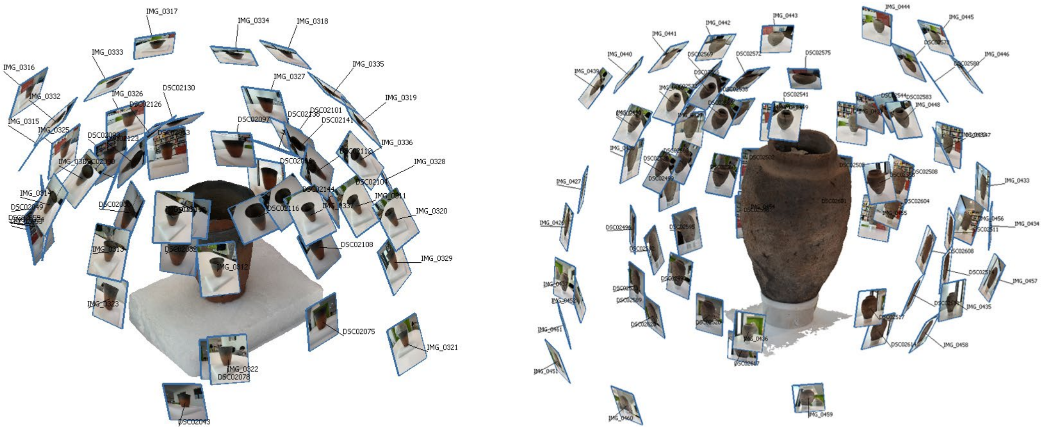

Fig. 1. Photogrammetric acquisition, Metashape interface. On the left a Class B vase, Black Topper Pottery; On the right a Class R Rough Faced vase.

The proprietary HandyScan software was used to manage the scanner information, the Agisoft Metashape software for the calculation of the photogrammetric cloud, the first mesh and final texture (Fig. 1) and the 3DReshaper software for cloud cleaning and mesh refinement operations. The output of these operations is a high fidelity textured mesh of the objects under examination (for the pair of Class B pots, Black Topper Pottery there are two meshes of 27,506 vertices and 29,113 vertices, for the Class R Rough Faced pots there are two mesh of 25,232 vertices and 28,849 vertices).

Dataset Creation

As suggested by the bibliography and the experiments conducted previously, a good synthetic dataset must be characterized by a good level of heterogeneity and contain sufficient images to be able to correctly recognize the instance to be classified. The proposed workflow for creating these render images for training AI frameworks is divided into 4 phases:
– The first phase consists in the generation of the textured mesh. The optimized, textured mesh is used as input to an algorithm written in Rhino's Visual Programming Language (VPL) environment. The mesh is recognized dimensionally through the creation of a bounding box which also identifies its barycentric point. This point is used as a base point for creating a sphere with a radius sufficient to contain the object. A chamber is positioned on each vertex of the tessellated sphere whose center of gravity is the center of gravity of the object. Depending on the different needs, the sphere is tessellated more or less finely: a smaller number of faces allows a reduction in calculation times, on the contrary a greater number of faces will generate a richer dataset. In this specific case, 320 photorealistic images of the single object were generated in two lighting conditions (to simulate two different locations within the museum environment). The images are 480 * 480 pixels in size and have been exported in .png format with no background. The automated process has integrated a PBR (V-Ray) rendering engine for better rendering (Fig. 2).
– Using a Python environment, the images containing the object's renderings are then processed to calculate its outline, to be used in the dataset annotation phase, and their insertion within images containing background scenes.
In the script we have developed for this use case, the images are first imported using the OpenCV library [Devjyoti 2021]. During the import process the alpha channel is preserved which allows us to exploit transparency to our advantage, using the border of the non-transparent region as the object contour.

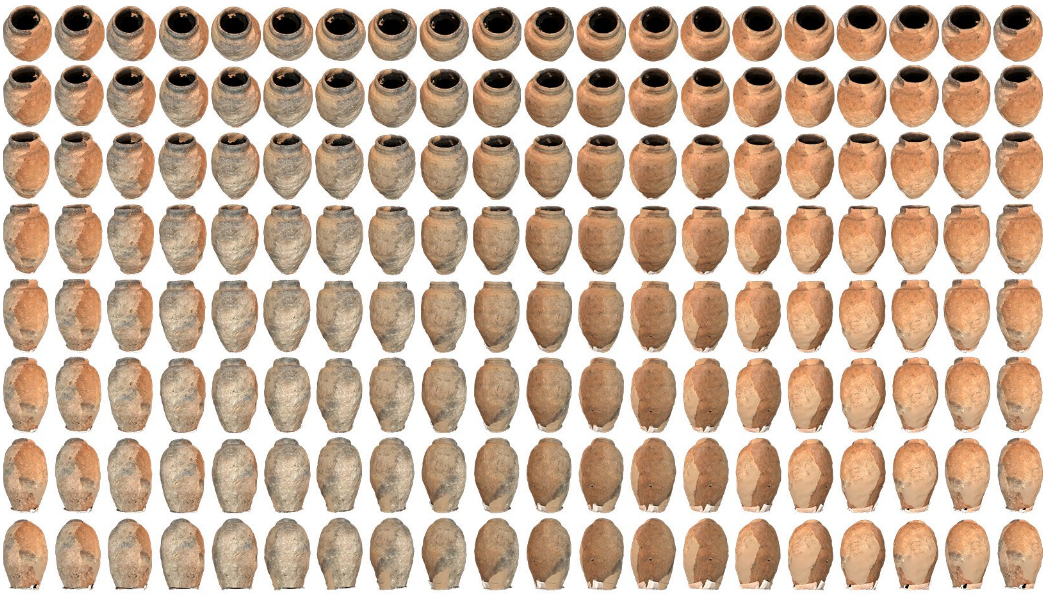

Fig. 2. Rendered images of a Class R Rough Faced vase.

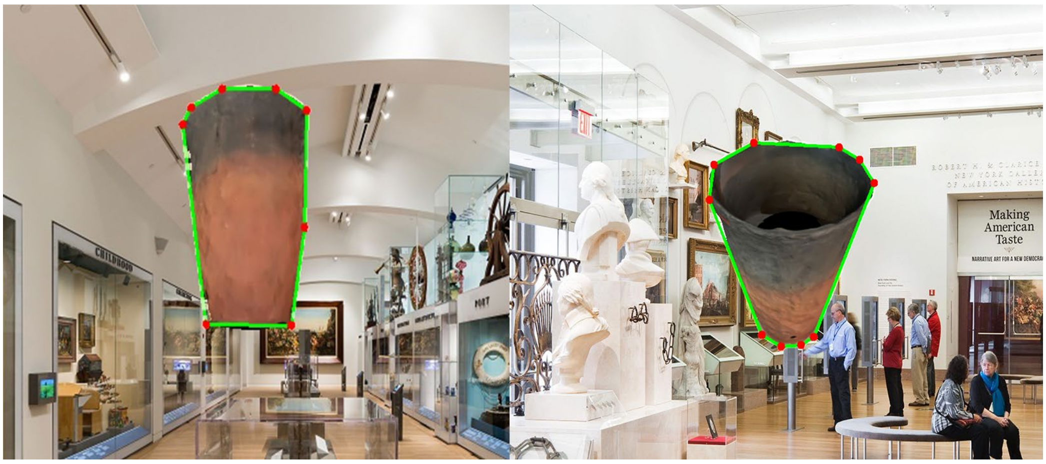

Fig. 3. Two images of the synthetic dataset. Random background of the SUN dataset and green outline annotation.

However, the contour identified with this method is too complex and jagged to be used for training our neural network. We then proceed by applying the Ramer-Douglas-Peucker algorithm, an algorithm that reduces a complex curve composed of numerous line segments into a similar curve but with fewer points.
– Once the outline has been extracted, the renderings are inserted into images containing backgrounds, which are useful for mitigating overfitting problems. The background images were extracted from the Scene UNderstanding (SUN) dataset [Xiao et al. 2010]. This dataset is commonly used to perform environment recognition operations. In this case we have extracted via API a set of random images, of cardinality equal to the dataset we generated, starting from the dataset in question. Our goal is to generalize as much as possible the environment in which we are going to insert our objects to ensure that the learning process of the neural network focuses on the objects and not on the background. To further increase the learning robustness, the images of the objects are anchored to a point of random coordinates within the background image, a scaling and rotation operation is also applied in advance to the object before the end of the operation. insertion. The parameters of these transformations are saved and then applied to the outline of the previously extracted object to ensure that it is consistent with the coordinate system of the background image in which the object was inserted.
– Once the operation is complete, the image is then saved together with the corresponding annotation file containing information on the outline of the object within the image. The annotation file is a .json format file compatible with the standard used by the VGG Image Annotator (VIA) tool. This format was chosen to maintain compatibility with manually annotated files.
The examples in Fig. 3 show two examples of annotated files with the outline (green) and its points (red) highlighted.

Neural Network Description

To carry out the recognition of the objects to be classified, two approaches based on different architectures have been explored. The two architectures considered are YOLOv5 and Mask_RCNN.
The most obvious difference between the two architectures is the method by which the searched object is highlighted within the image. The goal of YOLO is, being designed for real time detection scenarios, to identify the object as quickly as possible, yet only drawing a bounding box around it. On the contrary, the Mask_RCNN architecture [Waleed 2017], considered as the successor of Faster R-CNN, is a framework that allows you to perform object instance segmentation operations albeit with a slight performance impact [Buric et al. 2018].

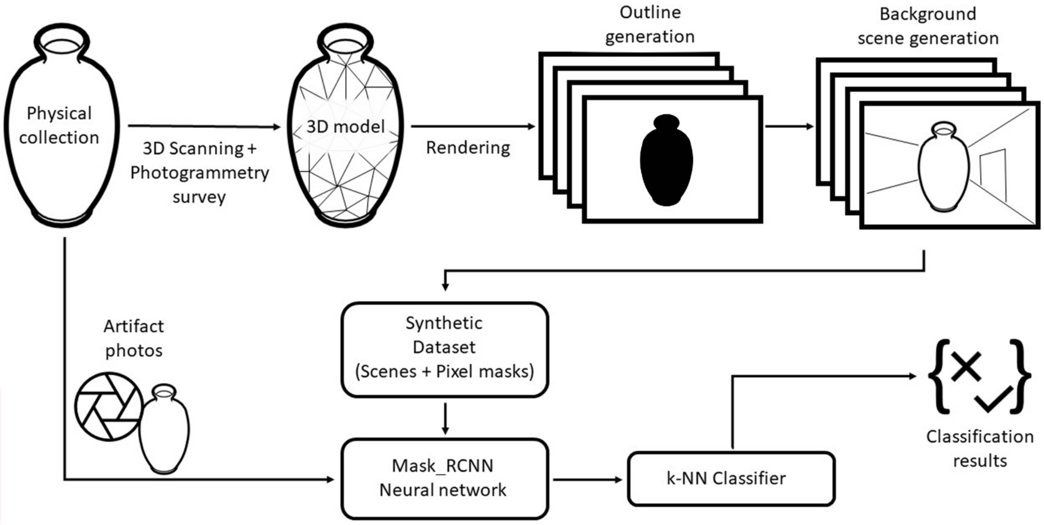

Fig. 4. Workflow overview.

Given the absence of stringent requirements on the time necessary to carry out the recognition of the objects in this use case, after an initial phase of experimentation it was decided to use the Mask_RCNN framework given the advantages brought by the semantic segmentation in the identification process of the artifacts treated by this paper.

Semantic segmentation, unlike simple object detection, provides more information. With this technique the object is not localized with a simple bounding box, the semantic segmentation provides the exact number of pixels and their position within the identified object as a result. In the event that there are more than one object in an image, each object is marked with its own color, allowing the individual objects to be identified visually. The possibility of extracting the objects from the background using the mask will allow us to use simple classification algorithms which won't be confused by the noise coming from the background data (Fig. 4).

The Mask_RCNN implementation used in this paper is available at Matterport / Mask_RCNN. Given the consistent hardware requirements to train the network from scratch, it was decided to use the transfer learning technique described above to reduce these requirements. In this case weights trained on the COCO dataset were used. The training images correspond to the ones generated by us synthetically. For this procedure, the images have been resized to have a maximum size of 128px on the long side. For the validation dataset, real images used to prepare the 3D models of one vase per pair were used, annotated manually. The images of the other vase of each pair were used as the testing dataset. As for the parameters of the neural network, the batch size was kept at 8, having a single Nvidia 1080Ti GPU available. Finally, a class called "vase" has been defined, corresponding to the objects we are going to recognize. In this first experimental approach it was decided not to carry out the classification of the vessels through Deep Learning but only subsequently with simple Machine Learning algorithms.

Fig. 5 shows the results of the detection process on images belonging to both types of vessels. You can see how the vases have been recognized with an average accuracy of about 90% and how the mask mostly respects the original shape of the objects.

As introduced above, in this specific case, to implement the last step of the recognition pipeline proposed by us it was sufficient to apply a clustering algorithm to the RGB values of the pixels contained in the regions identified by the Mask_RCNN network. The fact that the categories of vases used in this research showed some clear differences between them allowed us to classify them simply by using the color value of the pixels belonging to them. In fact, since the network allows exporting the points of the mask identified, these points have been used and imported by a further Python script which clusters their colors. The three dominant clusters of colors were indeed effective when used to classify the images, using a k-NN algorithm implemented in the Scikit-learn Python library. A future possible

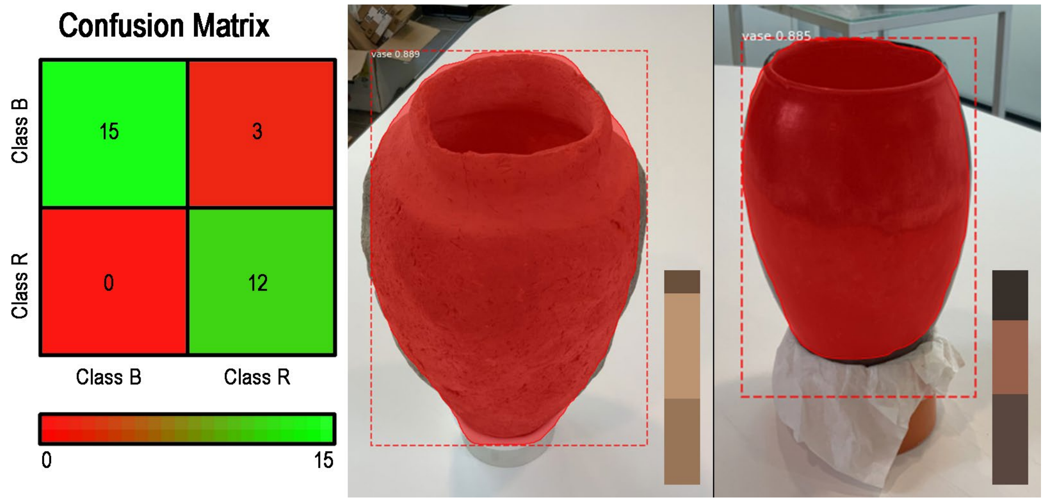

Fig. 5. On the left the Confusion Matrix sows how the accuracy of the k-NN classification algorithm on the test dataset (30 images, 15 for each class of vase) has been 90%, split into 100% for Class B vases and 80% for class R vases;
On the right the results of the detection process.

development of particular interest would be to evaluate the accuracy of a Mask_RCNN network in which the different vessels are considered as different classes, thus using only the DL approach to carry out the classification. This could be useful in classificating objects which are harder to distinguish.

Conclusion

One of the main limitations of Artificial Intelligence algorithms is their need to have more and more data for the training phases and in fields of application such as that of museum collections this problem is even more evident. This paper presents an approach to improve the accuracy of the detection of objects belonging to museum collections through the structuring of synthetic datasets automatically labeled with low budget instrumentation. Given the really promising results, the research group intends to further optimize the proposed workflow to make it more flexible and further scalable.

The application cases on which innovation can be implemented by implementing image recognition as described in the paper are many and can involve different users: from gamification applications for the general public; the creation of research tools for museum professionals.

In the first case, it is not difficult to imagine the programming of tools that improve the museum experience. In the bibliography it is known how AR systems, if properly combined with object detection algorithms, can expand the level of knowledge that can be accessed [Spallone et al. 2020]. Through these tools, the user, no longer bound to QR codes or didactic panels, can explore the collection and dwell on the artifacts that most intrigue him.

In the second case, through a different commitment of resources, one could arrive at tools for monitoring the collections present in the museum, or at the creation of *ad hoc* tools that simplify the operations carried out by the researchers to classify the new finds.

The working group undertakes to further study the topic and to identify case studies on which to apply the AI models trained through this workflow.

Notes

[1] It is interesting to report the conclusions of Nowruzi's research group as they noted that within a synthetic dataset it is more important to achieve a certain level of heterogeneity rather than a high photorealistic accuracy.

[2] Classification: By classification we mean the task of assigning a single label to a data (an image in our case) entering the model; Semantic segmentation: The segmentation of images is a fundamental part for Computer Vision systems which consists in partitioning an image into different regions representing the different objects; Object Detection: Detection is defined as the task of finding rectangular regions of an image, in which objects of interest are represented. These regions, called Bounding Boxes, are then classified to describe the object they contain; Instance Segmentation: Instance Segmentation is challenging because it requires the correct detection of all objects in an image and their segmentation or the definition of the exact perimeter and area occupied by the object. It therefore combines elements from classic computer vision tasks such as Object Detection and semantic segmentation.

[3] Among the various metrics calculated during the training of a neural network is the loss function, which allows you to have a measure of how much the algorithm has managed to learn from the data that has been provided. The error calculated by this function can have different components. One component is formed by the so-called irreducible errors, which cannot be reduced regardless of the algorithm that has been applied. Another factor is made up of the so-called reducible errors including, for example, the error defined by the difference between the value that was predicted by the model and the one you are trying to predict. When the value of this error is very high it is then possible that the model is experiencing an underfitting problem. When, on the other hand, it happens that the model has difficulty in identifying the correct predictions on real data and on the contrary the error is very small on the dataset used to train it, in this case we speak of a possible overfitting.

References

Buric Matija, Pobar Miran, Ivasic-Kos Marina (2018). Ball detection using YOLO and Mask R-CNN. In *2018 International Conference on Computational Science and Computational Intelligence*. Las Vegas: IEEE, pp. 319-323.

Cardenal Escarcena Francisco Javier, Mata Emilio, Castro P., Delgado Jorge, Hernandez M. A., Perez Jose Luis, Ramos M., Torres Manuel (2004). Evaluation of a digital non metric camera (Canon D30) for the photogrammetric recording of historical buildings. In *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 35 (B5), 2004, pp. 564-569.

Devjyoti Chakraborty (2021) OpenCV Contour Approximation – PyImageSearch.
pyimagesearch.com/2021/10/06/opencv-contour-approximation/ (November 2020).

Handa Ankur, Patrauceanc Viorica, Badrinarayanan Vijay, Stent Simon, Cipolla Roberto. (2015) Synthcam3d: Semantic understanding with synthetic indoor scenes. In *arXiv preprint*, 2015, 1505.00171.

Hinterstoisser Stefan, Pauly Olivier, Heibel Hauke, Marek Martina, Bokeloh Martin (2019). An annotation saved is an annotation earned: Using fully synthetic training for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*. Piscataway: IEEE, pp. 2787-2796.

Lin Tsung-Yi, Maire Michael, Belongie Serge, Hays James, Perona Pietro, Ramanan Deva, Dollar Piotr, Zitnick Lawrence C. (2014). Microsoft COCO: Common Objects. In Fleet David, Pajdla Tomas, Schiele Bernt, Tuytelaars Tinne (eds.). *Eur. Conf. on Computer Vision*. Cham: Springer, pp. 740-755.

Marinescu Maria Cristina, Reshetnikov Artem, López Joaquim Moré (2020). Improving object detection in paintings based on time contexts. In *2020 International Conference on Data Mining Workshops*. Sorrento: IEEE, pp. 926-932.

Mitash Chaitanya, Bekris Kostas, Boularias Abdeslam (2017). A self-supervised learning system for object detection using physics simulation and multi-view pose estimation. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems*. Vancouver: IEEE, pp. 545-551.

Nowruzi Farzan Erlik, Kapoor Prince, Kolhatkar Dhanvin, Hassanat Fahed Al., Laganiere Robert, Rebut Julien. (2019). How much real data do we actually need: Analyzing object detection performance using synthetic and real data. In *arXiv preprint arXiv*, 2019, 1907.07061.

Papon Jeremie, Schoeler Markus (2015). Semantic pose using deep networks trained on synthetic RGB-D. In *Int. Conf. on Computer Vision*, 2015, pp. 774-782.

Rajpura Param, Bojinov Hristo, Hegde Ravi (2017). Object detection using deep cnns trained on synthetic images. In *arXiv preprint arXiv*, 2017, 1706.06782.

Spallone Roberta, Palma Valerio (2020). Intelligenza artificiale e realtà aumentata per la condivisione del patrimonio culturale. In *Bollettino SIFET*, 2, 2020, pp. 1-8.

Tremblay Jonathan, To Thang, Birchfield Stan (2018). Falling things: A synthetic dataset for 3d object detection and pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. Piscataway: IEEE, pp. 2038-2041.

Tomalini Andrea, Pristeri Edoardo, Bergamasco Letizia (2021). Photogrammetric survey for a fast construction of synthetic dataset. In Giordano Andrea, Russo Michele, Spallone Roberta (eds.). *Representation Challenges. Augmented Reality and Artificial Intelligence in Cultural Heritage and Innovative Design Domain*. Milano: FrancoAngeli, pp. 215-219.

Waleed Abdulla (2017), Mask R-CNN for Object Detection and Segmentation.
github.com/matterport/Mask_RCNN (November 2020).

Xiao Jianxiong, Hays James, Ehinger Krista, Oliva Aude, Torralba Antonio (2010). Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*. Piscataway: IEEE, pp. 3485-3492.

Authors
Andrea Tomalini, Dept. of Architecture and Design, Politecnico di Torino, andrea.tomalini@polito.it
Edoardo Pristeri, Leading Innovation & Knowledge for Society, LINKS Foundation, edoardo.pristeri@linksfoundation.com

Isbn 9788835141945