Decay detection in historic buildings through image-based deep learning. Silvana Bruno, Rosella Alessia Galantucci, Antonella Musicco

Decay detection in historic buildings through image-based deep learning

Silvana Bruno, Rosella Alessia Galantucci and Antonella Musicco

DICATECh, Department of Civil, Environmental, Land, Construction and Chemistry, Politechnic University of Bari.
Email: silvana.bruno@poliba.it

Abstract: Nowadays, built heritage condition assessment is realized through on-site or photo-aided visual inspections, reporting pathologies manually on drawings, photographs, notes. The knowledge of the state of conservation goes through subjective and time or cost consuming procedures. This is even relevant for a historic building characterized by geometrical and morphological complexity and huge extension, or at risk of collapse. In this context, advancements in the field of Computer Vision and Artificial Intelligence provide an opportunity to address these criticalities. The proposed methodology is based on a Mask R-CNN model, for the detection of decay morphologies on built heritages, and, particularly on historic buildings. The experimentation has been carried out and validated on a highly heterogeneous dataset of images of historic buildings, representative of the regional Architectural Heritage, such as: castles, monasteries, noble buildings, rural buildings. The outcomes highlighted the significance of this remote, non-invasive inspection technique, in support of the technicians in the preliminary knowledge of the building state of conservation, and, most of all, in the decay mapping of some particular classes of alterations (moist area, biological colonization).

Keywords: built heritage; historic buildings; decay detection; deep learning; Mask R-CNN.

Cite as: Bruno, S., Galantucci, R.A., and Musicco, A. 2023. 'Decay detection in historic buildings through image-based deep learning'. *VITRUVIO - International Journal of Architectural Technology and Sustainability*, *8*(special issue), pp. 6-17. https://doi.org/10.4995/vitruvio-ijats.2023.18662

1. Research background

Currently, building surveyors working in refurbishment and maintenance activities execute the condition assessment during on-site or photo-aided visual inspections, reporting pathologies manually on drawings, photographs, notes. Nevertheless, a more immediate method could be advantageous, to assess the conservation state in a more effective and timely manner. This is even relevant for an historic building characterized by geometrical and morphological complexity and huge extension, or at risk of collapse. To address these issues, a real-time detection of building pathologies from on-site images can support experts in the process of building diagnostic. In this regard, advancements and innovative findings in Computer Vision (CV) represent a beneficial opportunity in decay recognition, related to several operative fields (road, bridges or buildings' inspection). CV comprehends digital image processing methods, in which manually annotated features are used to create rules to recognize the selected features. In particular, these methods have been developed to automatically identify features/objects within a set of images, and to continuously optimize themselves, for the detection of multiple or individual objects in images with a valuable accuracy. In the first place, Machine Learning (ML) handled these objectives, but several issues emerged from their application. Nowadays, Deep Learning (DL) models achieve a more efficient object recognition and object classification, with a good resulting accuracy. Indeed, it does not require a pre-processing step of the input raw data (e.g. image pixels. Among DL algorithms, Convolutional Neural Networks (CNN) proved to be particularly proficient, to solve issues connected to the information retrieval from images, such as classification, object detection, localization and segmentation. The learning process typically follows a multi-layered architecture, constituted by neural networks that analyse the most relevant features. These layers are sliding 2D windows within images. For the sake of clarity, object detection consists of image classification and object localization (regression), to recognize specific objects. In particular, image classification predicts the class of a single element inside the image through a probability distribution vector generated by the output layer over the pre-defined classes. Object localization entails the identification of the position of objects inside an image through bounding boxes. CNN processes images, characterizing each feature with a matrix of values, to be determined and compressed into 3D tensors. Then, transposing the image, tensors generate a series of convolutional and pooling layers, able to extract data from the image segment and to compress it into a representative structured array. This is an iterative pipeline, which ends with the formation of predictions. In order to achieve an effective object detection, segmentation can be added to divide an image into subgroups/pixels and create regions. In this way, segments with no information are not considered and the entire process is streamlined. Image segmentation creates a pixel-wise mask for each object, highlighting foreground elements. Thus, the assignment of labels to pixels facilitates the creation of boundaries of objects. Segmentation can pursue different aims. Semantic segmentation partializes images, including similar objects and uniformly coloured in a general category (i.e. damage). All the individual objects are identified via instance segmentation, so as each object class has its own colour. The instance segmentation involves object detection, object localization and object classification (He *et al.*, 2017).

In AECO (Architecture, Engineering, Construction and Operation) domain, in the context of building defect detection, some attempts have been done, using CNN to create tools for fast and effective survey and communication of building conditions, in order to timely intervene with repairs and maintenance. Several studies have been carried out, to develop crack detection approaches, mostly for road condition assessment. In the past, crack recognition was based on digital image processing, with the manual identification of relevant features for the recognition. The rapid development of computer hardware and the massive acquisition of images about road pavement's cracks lead to the implementation of Deep Learning systems, particularly based on convolutional neural networks. Some researchers added deep learning-based quantification tasks to the simple recognition, in order to quantify the defect width or length (Kim and Cho, 2019). Some others applied crack detection and quantification to photogrammetric reality-capture data (Kalfarisi *et al*., 2020; Wu *et al.*, 2020). Again for crack detection purposes, the U-net model, a fully convolutional network, turned out to be more robust, effective and accurate than Deep Convolutional Neural Networks (DCNN or CNN), with smaller training sets, also in heterogeneous conditions (extreme light, noisy background, thin cracks) (Liu *et al.*, 2019). The joint of pre-training and migration learning, both in Faster R-CNN (Regional Convolutional Neural Network) and Mask R-CNN (instance segmentation frameworks), allowed to solve the criticality of labour intensive training (Xu *et al.*, 2022). Perez et al. (2019) evaluated the use of CNN towards a real-time automated detection and localisation of key building defects, e.g., mould, deterioration, and stain, from images taken from recent built assets. The proposed model is based on pre-trained CNN classifier of VGG-16 (compared with ResNet-50 and Inception models), with class activation mapping (CAM) for object localisation (Perez *et al.*, 2019). In Cultural Heritage domain, efforts have been done to automatize the detection of visible decay on masonry surfaces, through the implementation of Machine

Learning. In particular, Mishra (2021) gave an extensive overview of ML methods for damage detection, diagnosis and monitoring, aimed at a proper retrofitting of historic heritages, highlighting the availability of limited data sets for ML applications in heritage buildings condition assessment (Mishra, 2021). In addition, the integration of non-destructive tests, simulation models and in situ studies should be investigated for ML-aided diagnosis. Hatir *et al*. (2020) compared ANN and CNN for weathering type recognition; the results' accuracy rates are similar (93.95% and 99.4%, respectively). However, CNN proved to be more effective and reliable, both in terms of time and classification outcomes. Nevertheless, this model is not able to identify multiple types of defects at once, and the entire image is detected as containing only one category (Hatir *et al*., 2020).

1.1. Research aim

Despite some research works have been carried out for decay detection in historic buildings, these are not extensive and requires in-depth investigation, especially in respect of multi-object detection and instance segmentation, with high accuracy and low hardware labour, but also with smaller input data sets for training and testing. For this reason, the paper aims to propose an innovative deep learning based-approach, the Mask R-CNN, for an expeditious detection of multiple kinds of decay morphologies, starting from 2D images, taken in different environmental conditions, image sensors and resolution (old photo-cameras, smartphones). In particular, the experimentation exploited a dataset deriving from a collection of on-site images of several historic buildings located in south Italy, diversified for epoch, materials and constructive techniques. Indeed, these buildings belong to different typological-morphological constructions: castles, monasteries, noble buildings, rural buildings. As far as the paper structure is concerned, after the introduction of the main context of the research and a general framework of the literature background (Section 1), the methodology is illustrated (Section 2). Section 3 concerns the technical specifications for the experimental application. Results and discussion are reported in Section 4, focusing on some performance indicators, to support the selection of the best training, in order to obtain a decay prediction with high accuracy and low computational cost. In the end, the experimental application allows the evaluation of this remote diagnosis technique, for a real-time identification of decay patterns in historic buildings (Section 5).

2. Materials and Methods

As previously introduced, the proposed methodology is based on a Mask R-CNN model, for the detection of decay morphologies on built heritages, and, particularly on historic buildings, as defined by the main sectorial standards (UNI, 2006; ICOMOS ISCS., 2008). The main goal is the automatic assessment and recognition of multiple alterations on images of buildings or their parts, in order to support the technicians within the diagnostic process. Mask R-CNN is a state-of-art object instance detection approach, which detect objects in an image and simultaneously generates high-quality segmentation masks for each instance (He *et al*., 2017). It entails an evolution of Faster R-CNN, which is, in turn, an extension of R-CNN (Region Based Convolutional neural network). R-CNN works creating bounding boxes, in correspondence of the regions of the objects to identify (Regions of Interest - RoI), which are then analysed with separate convolutional networks, for the association of plural image regions into a specific class. A first advancement is represented by Faster R-CNN, with its Region Proposal Network (RPN), which preliminarily proposes all the detectable objects into a single image. For the peculiar objectives of the research work, a further development of R-CNN has been considered: a mask implementation within the Faster R-CNN, producing an additional output branch, representing the object binary mask (corresponding to each RoI) in a pixel-to-pixel way, with respect to the class label and bounding box predicting branches, typical of the R-CNN architecture (He *et al.*, 2017; Xu *et al.*, 2022; Odemakinde, no date; Khandelwal, 2019). This pipeline allows to accomplish two fundamental problems: multi-class classification, which concerns the identification of the object in the image, and, in this case the peculiar damage; and regression, to understand where this damage is located. They can be translated into the task of object localization, providing a prediction of both the presence of the object in the image, and the identification of its boundaries.

The proposed methodology consists into three main sub-processes: training, testing and validation of the model. Therefore, the image dataset is subdivided into three parts belonging to as many sub-processes of the workflow: train (70%), validation (10%) and test (20%) images, respectively.

The execution of the network goes through some specific tools, both in terms of software and libraries needed for the execution of the model. In particular, the development of the model starts from Mask RCNN library, TensorFlow and Keras (*Mask R-CNN library*, no date; *TensorFlow*, no date; *Keras*, no date). The computer vision part was supported by OpenCV (*OpenCV*, no date). While image processing and annotations was realized with Scikit image (*Scikit image*, no date) and Json (*Json*, no date).

2.1. Object detection model

Going into detail, an object detection model has been realized for each alteration class to be searched (like moist area or biological colonization), by exploiting a state-of-art Mask R-CNN, with pre-trained weights of COCO (Common Object in Context) database for a first training phase. This dataset supplies large amounts of annotated images of common objects inserted in their natural context (Lin *et al.*, 2014).

The procedure is organized as follows: on the one hand, images are examined, and proposals of probable areas are generated; on the other hand, the proposals are classified producing bounding boxes and masks, in correspondence of the objects. As previously introduced, Mask R-CNN provides three kinds of output (class label, bounding box, object mask). The whole architecture is articulated into three main stages, which, on their side, can be furtherly divided into sub-steps (Fig. 1):

1. *Feature extraction (backbone):* the first stage consists in the application of a convolutional neural network, for the extraction of feature maps from the images. This operation entails a first low level feature extraction, which refers to edges or corners; and then a higher level of features, dealing with objects to be detected. The backbone step acts a conversion of the original image into its corresponding feature map, which constitutes the input for the following stages. In particular, Residual Network and Feature Pyramid Network (ResNet101+FPN) have been chosen as convolutional layers, because of their efficiency and celerity (He *et al.*, 2017). Residual Networks makes it possible to train hundreds of layers. Indeed, ResNet101 has an architecture of 101 layers (He *et al.*, 2016). While FPN, which is applied after ResNet 101, exploits a pyramidal structure to extract multiple features at different scales (Li *et al.*, 2019).

2. *Region Proposal Network (RPN):* the second stage of Region Proposal Network is a neural network that works proposing probable regions, which could likely contain objects of interest, with a sliding-window technique applied to the input feature maps. RPN identifies candidate anchor boxes on the feature map, according to two main parameters: scale and aspect ratio. Given the possible boxes, the network establishes whether they belong to background or foreground, and fits the last ones to the objects, by calculating an offset of the x,y coordinates (centre of the box), and width and length of the box (Sagar and Jain, 2018; Ren *et al.*, 2017).

3. *Region Convolutional Neural Network. (R-CNN):* the third stage concerns the implementation of the region-based convolutional neural network (R-CNN), to the outcomes of the first two stages (feature maps and probable regions), according to the following steps:

    a. *RoIAlign:* at this step, the foreground, fitted boxes, resulting from the RPN, are processed with RoIAlign, a pooling layer that converts non-uniform input size feature maps into fixed size output maps, by maintaining the exact spatial locations. It solves the misalignment between feature map and RoI on the original image. In particular, it computes values of the sampling points, through a bilinear interpolation of the nearby points on the feature map. This layer is one of the elements, which differentiates Mask R-CNN from Faster R-CNN, entailing instead a RoIPooling (He *et al.*, 2017).

    b. *Fully Convolutional Network (FCN)* the last part of the architecture enables the application of a Fully Convolutional Network, which is organized into two kinds of layers, according to the required output:

        i. *Fully Connected Layers -> Class and Bounding Box:* the first kind of layers entail the object detection and classification, which generates the final classes and the bounding boxes. They have the structure of a convolutional layer with depth of 1024 (He *et al.*, 2017).

        ii. *Fully Connected Layers -> Mask:* while the second typology of layers, supplementary to the Faster R-CNN model, concerns the mask retrieval, aimed at performing an instance segmentation. Each RoI corresponds to a unique object, to which a semantic segmentation is applied.

In order to evaluate the algorithm a loss function is considered, expressing the distance between the current output of the algorithm and the expected output. The loss function for Mask R-CNN consists in the combination of three different losses, in correspondence of the three outputs of the model:

$$L=L_{cls}+L_{box}+L_{max} \quad (1)$$

Where: L_{cls} (class loss) refers to problems of an improper classification of the detected object; L_{box} (bounding box loss) derives from an inaccurate localization of a correctly classified object; L_{mask} (mask loss) is related to the ground

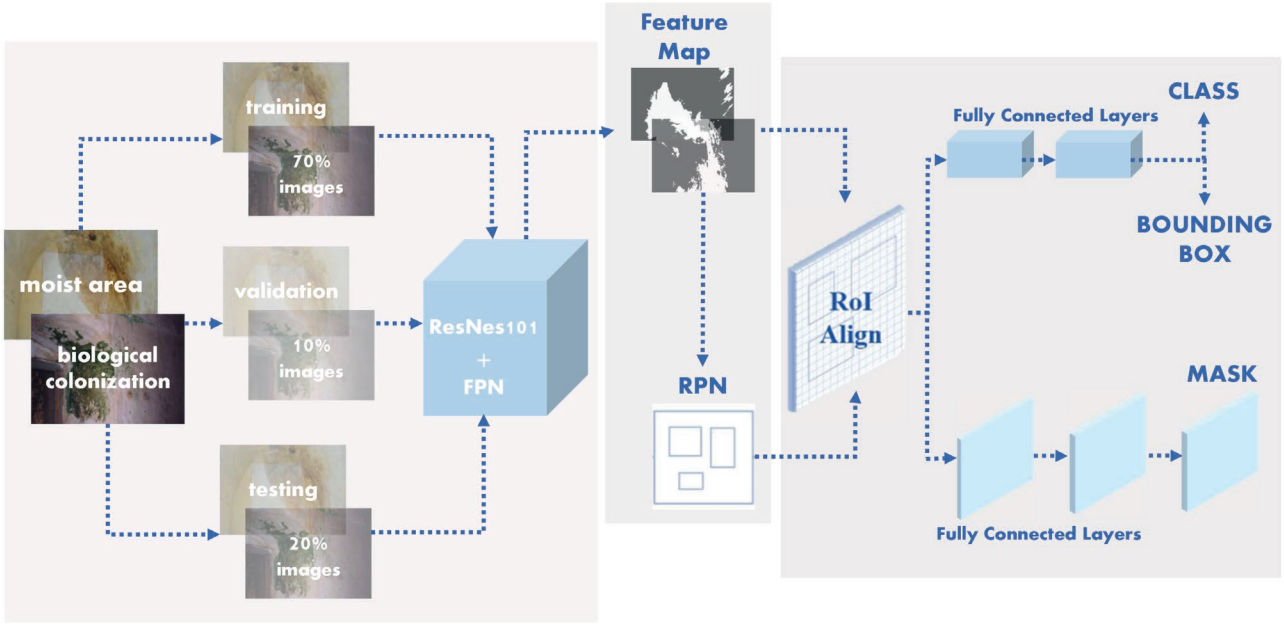

Figure 1 | Workflow of the model.

truth, and it is the average binary cross entropy loss (He *et al.*, 2017; Girshick, 2015).

A distinct loss function can be associated to the phases of training and validation. The training loss designates how the model fits the training data, while the validation loss appoints how the model fits the new data.

A further metric to be considered is the *Confidence*, a number from 0 to 1, expressing the percentage of confidence about the prediction of the algorithm. It is important to assign a confidence threshold, for the model training, in order to establish the limit of acceptance of a prediction.

2.2. Data preprocessing

In order to perform the model, it was necessary to prepare the dataset by resizing images that exceed the size of 1024×1024 pixels, because with larger dimensions it is not possible to feed the input images. A further passage concerns image annotation, which allows the machine to learn automatically, through the assignment of metadata and labels to the input digital images. The annotation has been used to organize the image dataset and to localize the objects of interest in each image. This step has been realized, for training, testing and validation of each class, using COCO dataset format. It is characterized by five annotation types: 1) object detection, which draws shapes/polygons around objects in an image; 2) keypoint detection, which simultaneously detect the object and its key points (invariant to the image transformations); 3) stuff segmentation, which is a segmentation of stuffs on the scene at pixel-level; 4) panoptic segmentation, which combines instance and semantic segmentation, the first for the detection and masking of objects and the second for the attribution of labels to the detected objects; and 5) image captioning, for the generation of textual description of an image. In the present research, annotations are stored using the JSON format, and, among the annotation types, the object detection mode has been selected. In particular, annotations and categories have been added to the images, which refers to bounding box information, segmentation, keypoint, and other label information for multiple tasks, but also to the list of categories, corresponding to the different classes of alteration.

2.3. Performance evaluation metrics

In order to evaluate the model performance, some main metrics have been considered, resulting from the combination of four kind of pixels' indicators: *True Positive (TP),* correctly predicted as positive; *True Negative (TN),* correctly predicted ad negative; *False Positive (FP),* falsely estimated as positive; *False Negative (FN),* falsely estimated as negative. In particular, the accuracy of the model is based on the following metrics:

- *Precision*, a measure of the prediction's accuracy, which means how many predictions are correctly

predicted as positive, with respect to the total predicted positive observations.

$$Precision=TP/(TP+FP) \quad (2)$$

- *Recall*, an indicator of how accurate the positive predictions have been, that is the ratio between True Positives and the actual positive predictions.

$$Recall=TP/(TP+FN) \quad (3)$$

- *Average precision* (AP), an indicator that combines the Precision (P) and Recall (R) curve into a single value, representing the average of all precisions.

$$AP=\sum_n[(R_n-R_{(n-1)})\times P_n] \quad (4)$$

- *Intersection over Union* (IoU), ratio between area of overlap and area of union, of ground truth and predicted segmentation.

$$IoU=(Area\ of\ Overlap)/(Area\ of\ Union) \quad (5)$$

- *Mean Average Precision* (mAP), a metric specifically used for object detection models like Mask-RCNN, calculated as the mean of the average precisions (APi) for all the classes (N).

$$mAP=(1/N)\times\sum_{i=1}^{N} AP_i \quad (6)$$

3. Experimental set-up

The experimental application concerned two kinds of damages, moist area and biological colonization (as defined by ICOMOS 2008 (ICOMOS ISCS, 2008), which are the easiest damage patterns to be recognized due to their peculiar morphological and chromatic features. The dataset is composed by RGB images, characterized by different dimensions, from a small size (748×1200 pixels) to a large size (3500×4000 pixels). In particular, 248 images have been considered for the moist area, and 142 images for biological colonization (Table 1). The large images have been resized to 1024×1024, because larger sizes are not consistent with the model's requirements.

Table 1 | Dataset of the two kinds of damages.

	Dataset	Training Set	Validation Set	Test Set
Moist area	248	173	49	26
Biological colonization	142	100	27	15

During the training of the model, a RoI is considered as positive if the IoU is equal or greater than 0.5, otherwise it is considered as negative. If RoI is positive, the mask loss can be defined. For the configuration of this model, two classes have been considered for each dataset, one for the object and another one for the background. In this case, the name of the objects corresponds to the name of the classes (moist area or biological colonization). If the confidence of the detection is less than 90%, the detection is skipped in the implementation. In addition, the number of training steps per epoch is defined as 64/100. The learning rate for the training is of 0.001. It is also necessary to specify the layers on which to perform the training. In this case, both the RPN and the mask convolutional layers have been included in the training phase. Two different numbers of epochs have been compared (50 and 100), in order to choose the one which produces lower training losses. After training the model, for each epoch size and for each class, a weight has been retrieved.

4. Results and discussion

The network has been performed for the two classes (moist area and biological colonization). The outputs of the model, for each image of the dataset, are represented by a bounding box, related to the object detection task, and a segmentation mask, to which the class is associated. They are confronted with the ground truth image.

To evaluate the model performance, three of metrics indicators have been used: Average Precision, Intersection over Union and Confidence. These parameters are necessary to understand whether the prediction's output is a true positive or false positive, because sometimes the classification is correct, while the bounding box location is wrong. Indeed, in order to obtain a true positive, three conditions must be verified: confidence greater than 0.9 (otherwise the model skips the image), AP between 0 and 1, and IoU higher than 0.5. Else, the output is a false positive.

For the moist area, the lower training loss results from 100 epochs (0.087), while for 50 epochs it is 0.43. Therefore, the 100 epochs weights can be considered as the most effective to predict the test images. The total time for the training has been around 15 hours and 30 minutes. For validation purposes, the mean Average Precision (mAP) has been calculated, in correspondence of two different sets of images (10 and 25). The mAP(10) is 0.4, which means that 4 images out of 10 have "1" as Average Precision (AP); whereas mAP (25) is 0.51, corresponding to 12 images out of 25, with AP equal to "1". Table 2 summarizes all the main metrics for training, validation and testing of the moist area. Afterwards, the model has been

Table 2 | Training (Tr), Validation (Va) and Testing (Te) calculated metrics for Moist area (for 173 images).

Number of epochs	Training Time	Training loss	Image size	Validation time	Mean average Precision (Va)	Testing Time	Mean average Precision (Te)
-	s	-	-	s	-	s	-
50	26,539	0.431	10	193.313	0.3	115.121	0,8
			25	542.684	0.5	282.086	0,8
			50	-	-	563.285	0,76
100	55,172	0.087	10	109.266	0.4	98.771	0,8
			25	293.099	0.5133	243.535	0,8
			50	-	-	382.000	0,74

tested with three different sizes of test images (10, 25, 50). The mean Average Precision on testing data for both the epochs 50 and 100 images is higher then 0.5. Thus, the moist area prediction is good (Table 2).

In Fig. 2, the results of moist area prediction are shown, for two different cases. The overall confidence for the object detection is 0.99. For all the images, the Average Precision is equal to 1, and the IoU is higher than 0.5

For the biological colonization, the weights of 100 epochs have been considered to predict unseen test images, because the corresponding training loss is lower than for 50 epochs (0.417 and 0.831, respectively). Also in this case, the model has been tested with different sizes of test images (13 and 27). The mean average precision, for the validation of a set of 13 images, is 0.34, which means that the prediction for biological colonization is not sufficient. The mean average precision for the testing with 100 epochs is low too (0.18) (Table 3).

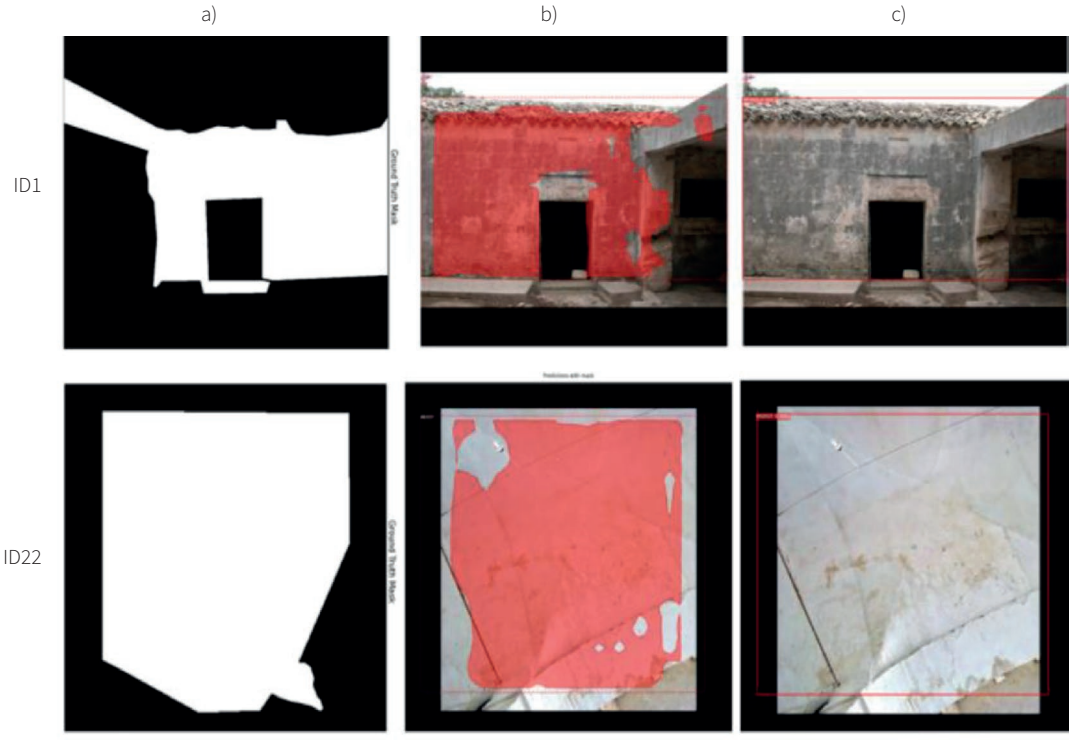

Figure 2 | a) ground truth mask (transposed image); b) segmentation mask; c) bounding.

Table 3 | Training (Tr), Validation (Va) and Testing (Te) calculated metrics for Biological colonization (for 100 images).

Number of epochs	Training Time	Training loss	Image size	Validation time	Mean average Precision (Va)	Testing Time	Mean average Precision (Te)
-	s	-	-	s	-	s	-
50	36014.45	0.831	13	183	0.26	157.059	0.17
			27	-	-	249.134	0.21
100	62942.75	0.417	13	109	0.34	138.001	0.26
			27	-	-	270.388	0.18

In Fig. 3, the results for three cases of biological colonization are shown. The overall confidence for the object detection is 0.98. While the Average Precision for the three images is 1, and the IoU is 0.77, 0.62 and 0.69, respectively.

5. Conclusions

Recently, improvements in Deep Learning, concerning the recognition of hidden and inexplicit recurrent patterns from raw input data, allows their implementation in various and diverse domains, from medicine to built

Figure 3 | a) ground truth mask (transposed image); b) segmentation mask; c) bounding box, for biological colonization

heritage assessment. Nevertheless, literature review highlighted the need for further investigations about reliable and time-effective procedures and methods to classify, detect and segment building decays. To this end, this work implements and validates a novel AI-driven system to detect and classify damages occurring on a building surface. In particular, the research scope involves the recognition of decay morphologies, like moist area and biological colonization. The proposed object detection and segmentation model exploits Mask R-CNN, to obtain an advanced instance segmentation and multi-object detection, employing a small input data set for training and test.

As far as moist area is concerned, the model provides highly accurate outcomes (mean average precision of the testing phase varying from 0,74 to 0,8), unlike the predictions for biological colonization, which are not as much accurate. Future developments could tackle these issues, by expanding the input data for the training of biological colonization images. In addition, this specific class could be divided into further subclasses, such as lichen, mousse, mould and plant, characterized by different features among each other (colour, shape, edges).

Anyway, it is important to underline that the model can provide an expeditious automated decay recognition and classification of two different kinds of damage, starting from heterogeneous quality, type or conditions image datasets. Therefore, it could represent a remote, non-invasive inspection technique, in support of the technicians in the preliminary knowledge and monitoring of the building state of conservation.

Furthermore, the test phase on new images takes place in maximum 10 min, which makes the implementation of the decay detection suitable and advantageous for the building condition assessment and aid decisions for conservation and maintenance activities.

Acknowledgments

The research has been developed under the supervision of Prof. Tommaso Di Noia (Department of Electrical and Information Engineering - DEI, Polytechnic University of Bari) and Prof. Fabio Fatiguso (Department of Civil, Environmental, Land, Construction and Chemistry – DICATECh, Polytechnic University of Bari). While Daniele Malitesta (PhD student, DEI) and Manish Chinthakindi (Master student, DEI) dealt with the implementation of the informatic aspects, related to the application of the Mask R-CNN to the experimental dataset.

References

Girshick, R. (2015). 'Fast R-CNN', *Proceedings of the IEEE International Conference on Computer Vision*, 2015 Inter, pp. 1440–1448. https://doi.org/10.1109/ICCV.2015.169

Hatir, M.E., Barstuğan, M., and İnce, İ. (2020). 'Deep learning-based weathering type recognition in historical stone monuments', *Journal of Cultural Heritage*, *45*, pp. 193–203. https://doi.org/10.1016/j.culher.2020.04.008

He, K., Zhang, X., Ren, S., and Sun, J.(2016). 'Deep residual learning for image recognition', *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-Decem, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). 'Mask R-CNN', in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988. https://doi.org/10.1109/ICCV.2017.322

ICOMOS ISCS. (2008). *Illustrated glossary on stone deterioration patterns.*

Json (no date). https://www.json.org/json-en.html

Kalfarisi, R., Wu, Z.Y., and Soh, K. (2020). 'Crack Detection and Segmentation Using Deep Learning with 3D Reality Mesh Model for Quantitative Assessment and Integrated Visualization', *Journal of Computing in Civil Engineering*, *34*(3), pp. 1–20. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000890

Khandelwal, R. (2019). *Computer vision: instance segmentation with mask R-CNN*. Dostupné z:" https://towardsdatascience.com/computer-vision-instancesegmentation-with-mask-r-cnn-7983502fcad1.

keras (no date). https://keras.io/

Kim, B., and Cho, S. (2019). 'Image-based concrete crack assessment using mask and region-based convolutional neural network', *Structural Control and Health Monitoring*, *26*(8), pp. 1–15. https://doi.org/10.1002/stc.2381

Li, X., Lai, T., Wang, S., Chen, Q., Yang, C., Chen, R., Lin, J., and Zheng, F. (2019). 'Weighted feature pyramid networks for object detection', *Proceedings - 2019 IEEE Intl Conf on Parallel and Distributed Processing with Applications, Big Data and Cloud Computing, Sustainable Computing and Communications, Social Computing and Networking, ISPA/BDCloud/SustainCom/SocialCom 2019*, pp. 1500–1504. https://doi.org/10.1109/ISPA-BDCloud-SustainCom-SocialCom48970.2019.00217

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C L. (2014). 'Microsoft COCO: Common objects in context', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8693 LNCS(PART 5), pp. 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

Liu, Z., Cao, Y., Wang, Y., and Wang, W. (2019). 'Computer vision-based concrete crack detection using U-net fully convolutional networks', *Automation in Construction*., *104*(January), pp. 129–139. https://doi.org/10.1016/j.autcon.2019.04.005

Mask R-CNN library (no date). https://github.com/matterport/Mask_RCNN

Mishra, M. (2021). 'Machine learning techniques for structural health monitoring of heritage buildings: A state-of- the-art review and case studies', *Journal of Cultural Heritage*, *47*, pp. 227–245.

Odemakinde, E. (no date) *Mask R-CNN: A Beginner's Guide*.

OpenCV (no date). https://opencv.org/

Perez, H., Tah, J.H.M. and Mosavi, A. (2019). 'Deep Learning for Detecting Building Defects using Convolutional Neural Networks', *Sensors*, *19*(16), p. 3556. https://doi.org/10.3390/s19163556

Ren, S., He, K., Girshick, R., and Sun, J. (2017) 'Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), pp. 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

Sagar, V., and Jain, S.J. (2018). 'Yield Estimation using faster R-CNN', *International Research Journal in Global Engineering and Sciences.*, *3*(1), pp. 110–116.

Scikit image (no date). https://scikit-image.org/

TensorFlow (no date). https://www.tensorflow.org

UNI (2006) 'UNI 11182 Beni culturali - Materiali lapidei naturali e artificiali - Descrizione della forma di alterazione - Termini e definizioni'.

Wu, Z.Y., Kalfarisi, R., Kouyoumdjian, F., and Taelman, C. (2020) 'Applying deep convolutional neural network with 3D reality mesh model for water tank crack detection and evaluation', *Urban Water Journal*, *17*(8), pp. 682–695. https://doi.org/10.1080/1573062X.2020.1758166

Xu, X., Zhao, M., Shi, P., Ren, R., He, X., Wei, X., and Yang, H. (2022) 'Crack Detection and Comparison Study Based on Faster R-CNN and Mask R-CNN', *Sensors*, *22*(3). https://doi.org/10.3390/s22031215