CLASSIFICATION AND OBJECT DETECTION FOR ARCHITECTURAL PATHOLOGY: PRACTICAL TESTS WITH TRAINING SET

Kai Zhang[1,*], Chiara Mea[1], Fausta Fiorillo[1], Francesco Fassi[1]

[1] 3D Survey Group, ABC Department, Politecnico di Milano, Via Ponzio 31, 20133 Milano, Italy
– (kai.zhang, fausta. fiorillo, francesco.fassi)@polimi.it, chiara.mea@mail.polimi.it

COMMISSION II

KEY WORDS: Deep Learning, Object Detection, Classification, Segmentation, Artificial Intelligence, Architectural Pathology.

ABSTRACT:
Image classification and object detection techniques have been widely discussed and developed in recent years; they are the basis of various prosperous applications, for example, real-time mapping. Promising as it is, the practical test in the cultural heritage field encountered multiple problems. In this paper, the authors attempt to share the research experimentations and the empirical knowledge focusing on the classification and detection of architectural pathology. The tests are built on elaborated training sets annotated with analysed and in-advance defined categories. The trained models were examined from the perspective of evaluation sets, model explanation and unseen datasets. The outcomes indicated the mistakes and confusions behind things and stuff in the object detection efforts, to which cultural heritage and architectural field are closely related. The model also reveals specific visual patterns for recognition from thousands of instances in the training set. By digging into different aspects of model performance, the potential and limitations of these techniques in practical applications can be better understood.

1. INTRODUCTION

Architectural pathology detection is an essential step for the maintenance of infrastructure, from modern architecture to cultural heritage to support restoration and preservation works. Examining and mapping extensive areas has been a time-consuming manual job even until now, generally done after a detailed survey of the spaces and synthesising the information on 2D mapped sections or orthophotos. Nowadays, the job of identification and mapping is entirely manual. Integrating artificial intelligence (AI) techniques is expected to accelerate this process, starting with automated annotation of images.
This means to solve the "What-is problem" automatically and quickly, and this is the solid ground upon which all the other data interpretation can be built, e.g. real-time mapping and semantic photogrammetry.
Recent years have seen fast development of AI. While deep learning (DL) methods for computer vision in conducting tasks like classification and object detection are becoming mature in many other fields, the application in the architectural field is rarely seen because they prepare the dataset, which is challenging due to the fact that CH is characteristic to map are too various and not standard. For this reason, preparing the correct training set needs a combination of 3D and 2D multiscale datasets.
However, if the provided training set is of high quality and matches the complexity of the model, the statistical methods can also be practical in this field. That typically needs large datasets and proper training. An example is given by LeCun et al. in 1989, who automated the recognition of digits by providing the dataset MINST and the well-known model LeNet. The dataset included 60,000 images of manually written digits training set for training and 10.000 for evaluation; each image contains 28*28 pixels in grayscale. Another famous example is given by Deng et al., 2009 attempting to use 1000 images to describe one synset.
Nowadays deep learning models are becoming more complicated by the years, capable of remembering description factors for thousands of categories. The models are enabled not only to recognize the presence of a certain objct on the image but also to locate it in the image itself. The application of those methods is limited because there are limited datasets annotated based on specific uses and practical scenes.
The main challenges of AI applications in the architecture field are mainly in the process of providing a profitable training set. Changes influence data acquisition in light, perspective and distances of tiny details, making categorising decay manually challenging.
This paper tests two deep learning tasks, classification and object detection, for detecting specific pathological patterns on architectural surfaces. It attempts to explain the results by discussing the problems that occurred during the preparation of the training set, examining the correlation of samples, and explaining the inference process.

2. RELATED WORKS

2.1 Introduction to the deep learning methods

Pathology detection is a challenging task because there is no clear boundary among the different types of pathology. This means that pathology can not be considered as "thing" (objects with a well-defined shape, e.g. cat, person) but must be considered as "stuff" (amorphous regions, e.g. sky, forest), following the definition of Caesar et al., 2018. It has addressed the issue of dataset preparation with 'thing' and 'stuff', emphasised the importance of stuff and discussed the contextual correlation to things.
In order to perform the pathology classification, the classical 3 steps are followed: Dataset preparation, model training and evaluation and explanation.
1) Dataset preparation has been discussed for a long (Deng et al., 2009; Everingham et al., 2010; Lin et al., 2015), including topics like the scale of the dataset (referring to the number of categories and instances), the semantic hierarchy of the classes, accuracy (reliability of the annotation), and diversity (appearance, positions, viewpoints). The data preparation

activities follow the suggestions from these previous experiences.
2) The model training models for image classification and objct detection in recent days are many. Among all, the most famous should be LeNet. Afterwards, typical man-crafted networks with limited depth of layer were developed, like VGG (Simonyan and Zisserman, 2015), and Inception Network (Szegedy et al., 2014). Residual networks such as ResNet were developed by He et al., 2015 which introduced the concept of residual connection, solving the problem of gradient vanishing and allowing layer depth increase.
Later Deep Learning models, like Faster RCNN (Ren et al., 2016) and YOLO (Redmon et al., 2016), are nowadays popular models used use for detection are they stand for two typical approaches for 2D object mapping. The RCNN, as a two-stage approach, starts with region proposals and then determines if objects are contained in each proposal. As a one-stage approach, Yolo directly asks the model to output box location and classes. A comparison of the two first-version models suggests that the YOLO is less sensitive towards small objects and less accurate in general, while faster in referencing.
3) To understand how the model is responsive to the data, one of the used methods is t-distributed Stochastic Neighbor Embedding (t-SNE) (Maaten and Hinton, 2008) that visualise the high-dimension data to examine the data and understand the model. It processes the data-points and locates them in a 2D map, exposing the similarities and confusion among the categories. Other techniques like class visualisation, Saliency maps (Simonyan et al., 2014), and Class Activation Map (CAM, Muhammad and Yeasin, 2020; Zhou et al., 2016) were later developed to understand the decisions made by the model. In this research, the choice is to use t-SNE and CAM.

2.2 Some related research

Multiple works have delved into the automatic attribution of semantic meaning to architectural data. Various AI approaches were adopted, triggered by the needs of individual case studies.
Some of them use the Machine Learning (ML) approach. This method will require a case-by-case training set. The outcomes are promising because the model is deliberately trained for the same case. This means that the 2D representation of samples possesses features that are more coherent to the describing categories; light environment and other surrounding conditions don't affect much. This is the case followed by (Grilli et al., 2018) that classifies 2D texture or orthoimages unwrapped from 3D models, projecting them onto 3D geometries for a better spatial understanding, performing a "texture-based" classification. Optimised models, orthoimages, and UV maps are created for each case under study.
Several works (Guerrieri and Parla, 2022; Kwon and Yu, 2019; Mishra et al., 2022; Pathak et al., 2021) have been seen using deep learning object detection to detect pathological issues and materials of the surfaces. The applications of those works are promising, but they deal with very specific objects and favourable conditions like high image quality and easily defined categories to be detected.
Currently, there hasn't been a general model or working approach to solve the problem in a more general way dealing with the detection of all the different types of surface deteriorations (related to stuff). This is real for pathology detection in 2D and even truer for a direct 3D approach.
3D object detection is another field under heated discussion. Deep learning models were trained to attribute semantic meaning directly to the point cloud. (Charles et al., 2017; Qi et al., 2018) Cultural heritage has been seen using machine learning and deep learning models to map architectural components (Malinverni et al., 2019; Pierdicca et al., 2020; Teruggi et al., 2020; Zhang et al., 2022; Zhang et al., 2022). While the pathology upon surfaces requires performing detection job on point cloud at high resolution, allowing calculations of geometric features at an extremely detailed level, this approach is not optimal for the reason of difficulty in data collection, heavy computational resources and relatively long-time processing.

3. METHODOLOGY

This paper mainly addresses the practical tests on architectural pathological issues for image classification and object detection tasks. Like the development history of deep learning technology, the statistic models were trained first to identify 'what' (classification) and then 'where' (detection).
The primal need is to figure out if the pathological issues are identifiable. The second is how they can be detected.
Then, the future task will be to use them for 3D mapping and other further utilities.

3.1 Designing the Categories

Families	Categories	Description
Biological colonisation	Biological colonisation (B.I.O.)	Colonisation of the material by plants and micro-organisms such as bacteria, cyanobacteria, algae, fungi, and lichen (symbioses of the latter three)
	Plant (P.L.T.)	Vegetal living being, having, when complete, root, stem, and leaves, though consisting sometimes only of a single leafy expansion (e.g., Tree, fern, herb).
Discoloration and deposit	Discoloration (C.H.R.)	Change of the surface colour in one to three of the colour parameters: hue, value and chroma.
	Crust and Deposit (CRU)	Generally, crust coherent to accumulation of materials on the surface. And deposit to accumulation of exogenic material of variable thickness.
	Subflorescence and efflorescence (S.N.E.)	Generally whitish, powdery, or whisker-like crystals on the surface. Subflorescences are usually hidden.
	Graffiti (G.R.A.)	Engraving, scratching, cutting or application of paint, ink, or similar matter on the surface.
Features induced by material loss	Alveolization (A.L.V.)	Formation, on the surface, of cavities (alveoles) which may be interconnected and may have variable shapes and sizes (generally centimetric, sometimes metric).
	Erosion (E.R.O.)	Loss of original surface, leading to smoothed shapes.
Crack and deformation	Crack (C.R.A.)	Individual fissure, clearly visible by the naked eye, resulting from separation of one part from another.
Detachment	Peeling (P.E.L.)	Shedding, coming off, or partial detachment of a superficial layer (thickness: submillimetric to millimetric) having the aspect of a film or coating which has been applied on the surface.
	Delamination (DEL)	It corresponds to a physical separation into one or several layers following the laminae.
	Disintegration (D.I.S.)	Detachment of single or aggregates of grains

Table 1. Families and used categories with their definitions.

One of the most essential steps of the process is to identify clearly the pathology categories under detection. The terminology in defining pathological issues comprises multiple types regarding various materials and causes. Furthermore, the visual aspect of the same deterioration type cannot be the same, being the result of a different combination of materiality, natural causes (climate and animals), building techniques, human interventions, etc.

The ICOMOS glossary for stone deterioration (Verges-Belmin and Stone, 2008), a widely recognised resource in the field of cultural heritage conservation, was used as the primary reference for implementing the semantic hierarchy and its accompanying categories. Please refer to Table 1 for further details.

To avoid confusion and simplify the categories to be used, some of the patterns that share similar visual representations have been grouped. Therefore, few categories are not included in the plane text: alga, lichen, moss, and mould for the biological colonisation; encrustation, film, glossy aspect, patina, soiling for the discolouration and deposit family; mechanical damage, microkarst, missing part, perforation, pitting for features induced by material loss; blistering, bursting, fragmentation scaling for detachment family.

Multiple categories mentioned in the ICOMOS glossary are not used for the training process. In some specific cases, ICOMOS degradation categories are rare to find samples, confusing or hard to recognise by both machines and humans. For example, the diagnosis of a deformation case requires knowledge of the original or earlier appearance of the object as a reference to determine if the current situation can be classified as deformation. For these reasons, these types of degradation are not considered at the moment but are left for further examination.

3.2 Data preparation

The training dataset was expected to deal with the problems addressed above, with the data mainly coming from field studies in Italy. This means the training set should have high similarities to the data acquired in aimed application scenes.

For this reason, the images used for the classification are taken almost right in front of, focused on the pathological area and showing only one type of decay. For object detection, the prepared photos were shot without strict rules regards the shooting position and image deformation. Moreover, some photos might include various object categories (see Figure 1).

This approach is aimed at solving two different practical tasks. The classification and segmentation tasks are expected to generate results from photos or images with better qualities, while object detection is expected to work in real-time, requiring robustness for more complicated situations.

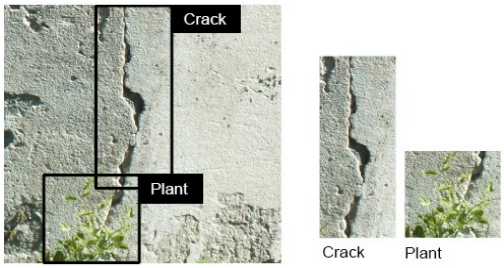

Figure 1. Samples for object detection and classification.

Photos for the training set have also been processed, especially to reduce the image size, which largely affects the training process, either leading to large memory consumption in CPU and GPU, or discarding objects that are relatively small for the annotation. The chosen main solution was to split some large photos into 9 parts. On one hand, it minimises the size of inputs; on the other, some irrelevant parts can be dropped to avoid redundant and negative training feeds.

3.3 Model training and evaluation

Considering the inference time, accuracy and model complexity, the chosen training models are ResNet18 (He et al., 2015) for the image classification and Yolo v5.

Residual networks allow skipping connections of layers (see Figure 2). By concatenating feature maps generated from the previous layer to the current, the network can continue processing without losing attention to details. This allows the model to avoid vanishing gradient and degradation problems, performing well as feature extractors in multiple other computer vision tasks.

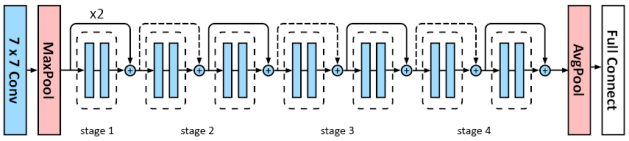

Figure 2. ResNet-18 Architecture.

Yolo architectures (see Figure 3) compared to the 'two-step' RCNN approach, are known for speed and convergence time while lowering precision. The used model architecture is comprised of mainly two parts: backbone and head. The backbone architecture is considered as the feature extractor; hence, it may vary from case to case, but commonly used is an optimised version of Darknet, namely CSP-Darknet53. This backbone typically includes four C3 blocks (CSP bottleneck blocks with 3 convolutions) eventually output through an SPPF (Spatial Pyramid Pooling – Fast) layer, with the first two C3 modules outputs catenated to the layers of detection head. The head in Yolo v5 accepts several outputs from the feature extractor; in the same logic of concatenation, the eventual detection head processes the outputs from the last three C3 modules, generating bounding boxes, confidence scores and class probabilities.

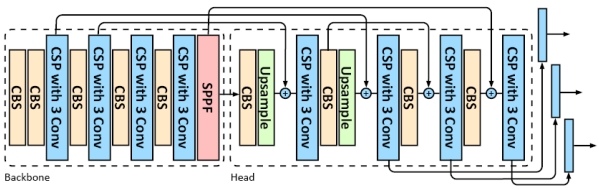

Figure 3. Yolo v5 architecture.

Part of the examination process is based on the numerical value calculated from the evaluation set that is randomly selected, taking up 10% of the overall annotated dataset.

These results suggest how well the model is fitted to the training set and, for similar scenarios, how the performances are in terms of precision. The representativity of the training and evaluating sets is namely their similarity with reality and it has also to be considered in the examination phase.

The model was also examined by using explanation tools, e.g. Class Activation Map (CAM). The explanation methods reveal how the model makes decisions by associating weights with each feature map in the final convolution layer. Summing the weighted feature maps, it allows visualising the triggering part of specific classes on a heatmap. The quality of the training model can be represented using the heatmap that can show in a

very intuitive way the areas that are closely related to the reference of human decision.

4. MODEL BEHAVIOR

4.1 Image classification

4.1.1 Dataset composition

4777 photos of different materials, including stone, ceramic, plaster, cement, and wood, were collected from multiple sources, including fieldwork, books, and the internet. After the pre-processing, the images were categorised into 12 pathological classes.

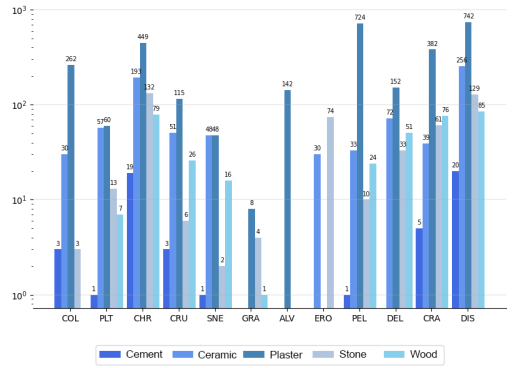

Figure 4. Histogram of the number of images per category

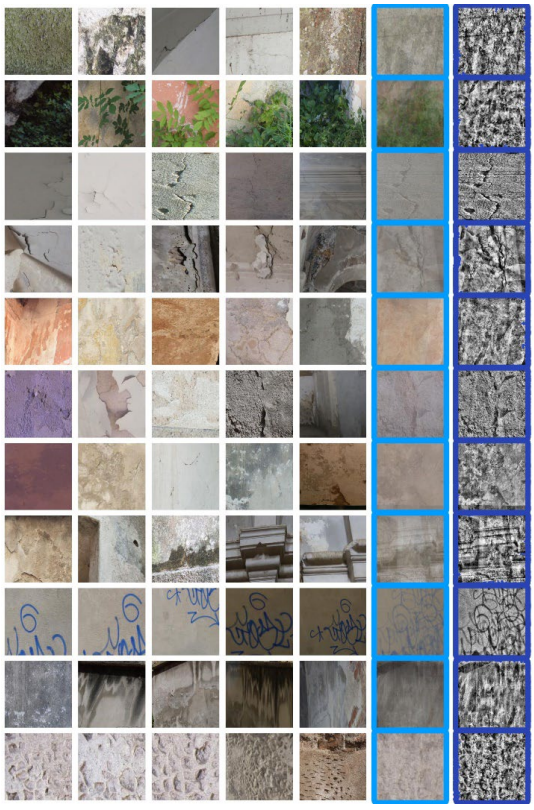

Figure 5. Example images from each category of material plaster, followed by an average image (marked by blue) and a strengthened pattern (marked by dark blue). From up to down list: Biological colonisation, Plant, Crack, Delamination, Disintegration, Peeling, Discoloration, Crust, Graffiti, Subflorescence & Effloresence and Alveolization.

Unfortunately, also in this case, as common in many examples, the dataset has a balance problem in terms of the number of samples. Figure 4 shows how many images are used for each category for the five different considered materials. In the future, as the annotation work proceeds, the constantly renewing dataset will enhance the model performances on categories for minor classes.

Taking the data of plaster, as an example, which was collected from the field precisely for the purpose of dataset preparation, the abundant samples and clear visual patterns enable proper training and discussion.

4.1.2 Training and evaluation

The training process started after the preparation of the training set. The model is trained on data related to plaster and stone material because they have more images (3500 images) and together cover all 12 classes.

Nowadays pre-trained model is commonly used for transfer learning. This could save time and computer resources because it leverages knowledge gained from conducting the source task on the target task. In the presented case the training set is ad hoc created and it is new and different from the most used as for example the ImageNet benchmark. For this reason, the ResNet-18 model is used and trained from scratch.

The model is trained with images processed into size 224*224, with 64 samples used in one iteration (batch size), with a learning rate (hyperparameter that determines the size of steps taken in the optimisation process) starting from 0.01 and with a scheduler that decays the rate each 30 epoch (a complete process of the entire training dataset) by a gamma equals to 0.1. The training process reached convergence after 800 epochs, taking 27.8 hours, using CUDA 12.2 upon Quadro P4000 8192MiB.

The first test achieved 49% accuracy on all the samples in the evaluation set. By removing non-nadiral photos, cropping the image to the wanted zone, and applying the trans-learning approach, the model acquired 80% accuracy with 500 epochs of training.

The best model achieved the 84% of accuracy. It got the optimal performances of 100% precision upon biological colonisation, graffiti and cracks, but some critical results with discoloration and disintegration with the precision of 66.7% and 42.8% accordingly.

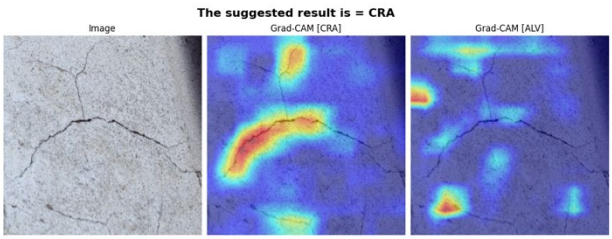

Figure 6. Gradient-CAM visualisation. For a photo (left) shot for target "CRA" (cracks), the middle and the right explain how the model made the decision for cracks and alveolization.

The class activation map (CAM) technique shows that the model is making decisions as expected. In fact is clear from Figure 6, that upon the image representing the crack, the activation of the indicated targets located the highest value upon the serpentine opening. If the activation map was asked to visualize the parts that activate other targets (the right heatmap in Figure 6 is CAM for alveolization) exposes the limits on image classification.

The model typically generates a single choice. To determine this decision, post-processing is performed on the output probability vectors. This involves applying a sigmoid function to map the

probabilities into a range of 0 to 1. The decision is then determined by selecting the maximum output probability.

As a matter of fact, it is not an ideal situation where the diagnosis of the pathology is made. The deterioration of the material is a process that lasts for a time, the causes and the effects represented on the surfaces are multiple. Photos shot for preservation purposes can record the complexity of pathologies. Aside from the primal and the most obvious pathology from the perspective of humans, there are secondary pathologies to examine. According to how many details are revealed, cracks, alveolization, erosion, and discoloration can represent themselves in various ways. Considering this complicated situation for detecting the pathology, the evaluation strategy that looks to merely numerical value is not sufficient.

The generalisation tests are conducted, but the results are not optimal. The model trained on the dataset of plaster and stone is tested on the dataset of other materials. It got precision on cement, ceramic and wood data of 0.2830, 0.3449 and 0.1562, with Top-3 accuracy equal to 0.5849, 0.6155 and 0.4265 accordingly. The model is also tested on a combined dataset that includes annotated images from all three materials (see Table 2). The confusion matrix suggests that crack, peeling, delamination, and discoloration confuse each other. The category plant acquired a high harmonic mean of precision and recall (F1-score) for the observably coarse pattern and high value in the green channel. Disintegration and discoloration also acquired relatively high scores. The model behaviour on the overall dataset (4777 samples) reaches 72% accuracy. This result is partially biased by the weight of the plaster and stone.

	Prec.	Recall	F1-score	Support
COL	0.1111	0.1515	0.1282	33
PLT	0.4143	0.4462	0.4296	65
CHR	0.3452	0.3677	0.3561	291
CRU	0.1200	0.0375	0.0571	80
SNE	0.2000	0.0308	0.0533	65
GRA	-	-	-	-
A.L.V.	-	-	-	-
ERO	0.0476	0.0667	0.0556	30
PEL	0.0522	0.1034	0.0694	58
DEL	0.0833	0.0163	0.0272	123
CRA	0.1789	0.1417	0.1571	120
DIS	0.3820	0.4931	0.4305	361
Accu.			0.2861	1226
Mac. Avg.	0.1612	0.1546	0.1471	1226
Wgt. Avg.	0.2671	0.2861	0.2667	1226

Table 2. Confusion matrix testing on unseen datasets, including cement, ceramic, and wood, using the model trained on plaster and stone.

4.2 Object detection

4.2.1 Dataset composition

Yolo v5 medium model is trained on the prepared manual annotated dataset. This dataset collected photo shots from in front and multiple other 'not standard' perspectives. It covers the samples of ceramic, plaster and stone to pursue a better balance of sample distribution. The dataset contains 1621 samples and 21008 instances, with an average of almost 13 instances in each image sample. 10 percent of the samples are randomly selected for evaluation.

Among all the categories, biological colonisation, peeling, and chromatic alteration are given the highest number of instances, above 3000 (see Figure 7). Graffiti on the other hand is found the least, although it's adequate to be defined by the provided samples.

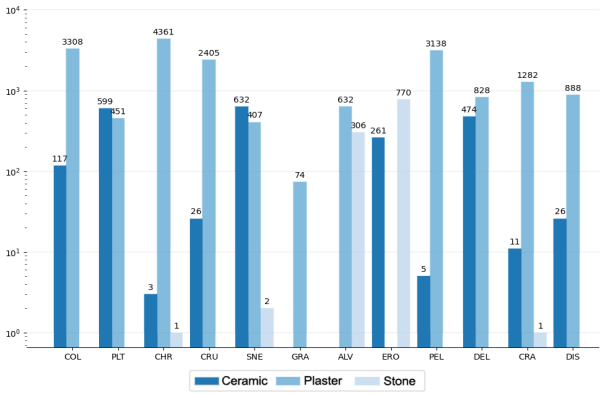

Figure 7. Instances distribution of the annotated dataset

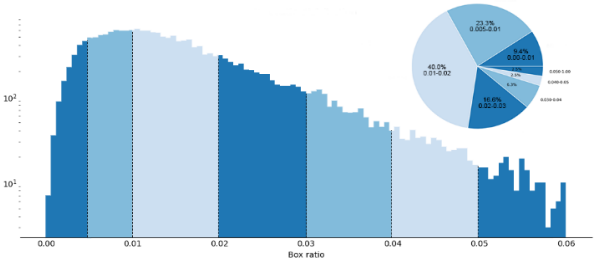

Figure 8. Box ratio distribution

The training set is prepared by marking bounding boxes which moderately include the object. A practical problem is highlighted: the boundary box has a square shape. Instead, the category to be correlated in many cases has an irregular shape not corresponding to a square area. Therefore, in many cases, the object category would occupy only a limited portion of the squared area. For this reason, it's decided to use continuous and adjacent boxes to mark the pathology, rather than a single box to include the whole area at once. If divide the area of the bounding box by that of the image, most of the used bounding boxes in the dataset are within the range of 0.005 to 0.03, bounding boxes under 0.1 take up 97.5% (see Figure 8).

4.2.2 Training and evaluation

For better management of files and computer memory, images are grouped as patches containing 200 images each. After the patches of samples were ready, the data was processed and fed to the YOLOv5m model. The training started from the pre-trained model, with the CSP-Darknet53 backbone layers frozen. The model is trained with image size 640*640 with batch size 16. The Yolov5 model used separately 3 learning rates for weight, bias and batch normalisation. The model uses 3 epochs to warmup, allowing each learning rate to reach 0.01. Afterwards, it uses a learning rate factor of 0.01 for the linear decay. The model reaches convergence after 2.08 hours of training, at around 80 epochs, using CUDA 12.2 upon Quadro P4000 8192MiB.

The trained model achieved acceptable results. The model needs 23.9ms inference time for each image. The bounding boxes plotted on the images take up 1% to 10% partition of the whole image. The best performances are the detection of biological colonisation, plant, erosion, graffiti, peeling, crack and disintegration. Others like crust, subflorescence and efflorescence, and alveolization in many cases turn out to be mispredicted.

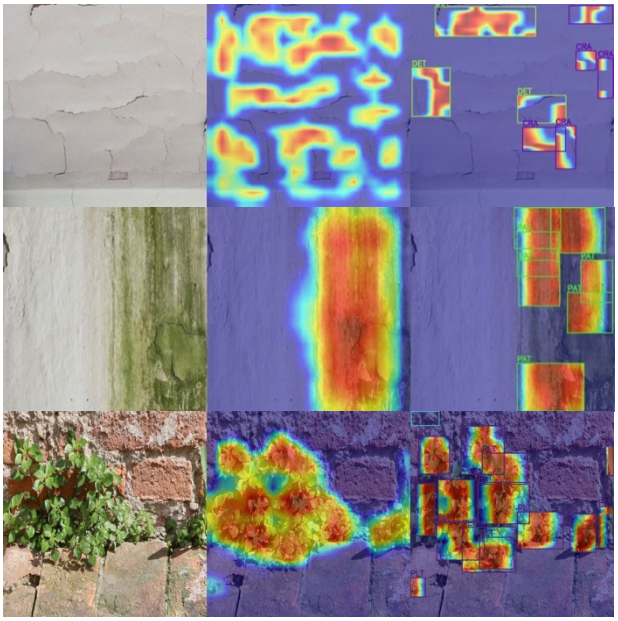

Figure 9. Eigen C.A.M. visualising activation map on the 8th layer for detecting peeling and cracks (up) biological colonisation (middle) and plant (down)

The model behaviour is better understood by using Eigen C.A.M. to plot the activation map upon the original image. Examples in Figure 9 can be considered as where the model is looking at when making decisions. To be noticed that in order to detect peeling, the model turns to look at the surrounding area of the cracking gap, where the detached pieces cast slightly the shading effect. It is different from the expectation that it will mainly look at the serpentine openings. Other examples also show that the model is trained well for finding out where to look for the interested object, even with distractions from the surroundings: In the middle, cracks and vertical patterns on the plaster surface are thought to be confusing to biological colonisation caused by the leaking water; at the bottom, similar green colour of biological organisms on the brick surfaces are thought to be confusing to class plant.

5. DISCUSSION

The practical application of deep learning methods in this paper is expected to suit the needs of the architecture preservation scenario. Hence, very importance is given to the reliability of the final results, to quantify (precision), and to reason (explanation). The training set preparation is fundamental for the corresponding tasks, for both classification and object detection. Considering the training and evaluation set, and future test sets, the evaluation of the model behaviour reveals the correlation between the provided samples and the related categories.

Confusion of classification may be attributed to the definition of categories. If merely examining the samples of the same material, taking plaster as an example, the pathologies might share similar visual features on images, as they are determined with a reference to the relatively plain and complete surface. The concept of decay on a surface is relative and related to the concept of 'completeness' Conceptually the more the pathology is defined based on the comparison to normal case, not standing alone to define itself, the worse it might be defined by the model, namely more confusion. Therefore, from the outcome of image classification, the most confusing categories are always: crack, peeling, delamination, discoloration and disintegration. Numerically, samples from each category were easily confused within these 5. On the contrary, among all the tests, the 'BIO' (biological colonisation) and PLT' (Plant), in most of the cases, present the highest F1 score (0.79-0.94). The reason seems to be clear: they are all green, regardless of their residing area. Additionally, these 2 classes can be differentiated from each other by the patterns and sizes of the shade.

It would be more reasonable to evaluate the model behavior by using Top-K accuracy and examine it using explainable artificial intelligence methods. Grouping some visually confusing categories, or output the corresponding family of the category can make the results more practical, leaving the possibility of sub-decision for the expert considerations.

Using the t-SNE technique, the samples from all materials are coloured according to the prediction results, and mapped in 2D based on the similarities(see Figure 10). Though the algorithm is stochastic, the results reveal the clarity of the designed categories to the annotated samples, by mapping the possibility vectors from the model output for all samples in 2D. From the outcome, biological colonisation, crust & deposit should be very well defined by the training set and are well differentiable from the other categories. At the clusters' edges, the mixture of colours can be spotted. The region of delamination is interrupted by samples of disintegration. The region of peeling by samples of discoloration and cracks. The regions of plant, discoloration, disintegration and crack unlike other categories located closer to the center, while the main parts of higher purity are separable from the mixture.

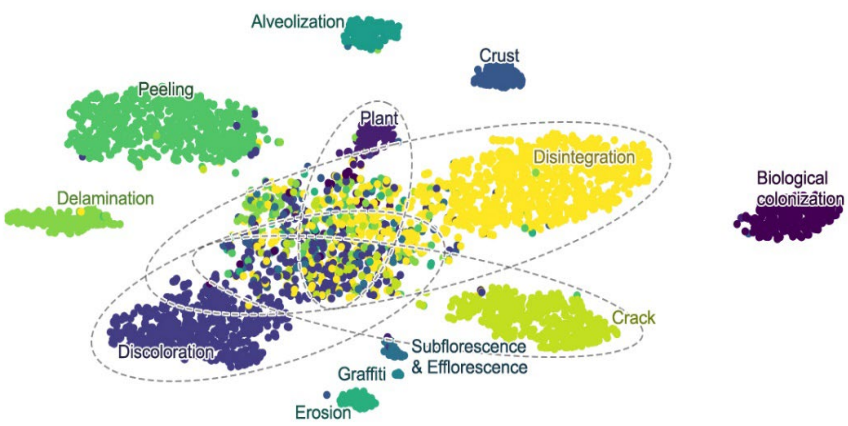

Figure 10. t-SNE visualisation across categories using the model trained on plaster and stone data, while mapping the overall dataset.

This t-SNE visualisation reveals mainly how categories for the unseen datasets (ceramic, cement and wood) can confuse with each other, considering the samples of plaster and stone are well fitted by the trained model. The mixture or the joint of clusters visually represents that those related classes are easily mistaken for each other. The mixture can indicate that the dataset annotation is under criticism. If the model is considered to be well-trained, then the problem is that image samples don't contain adequate for making the diagnosis of specific pathology, or the annotated category cannot fully represent the content.

The object detection model is good at detecting some specific pathologies, like biological colonisation, cracks, delamination etc. It allows exhaustive checks through all parts of the image. On one hand, this subjectivity is fast and helpful, the model will not be distracted by visual distortion, complicated patterns and light environment. On the other hand, the model will give unwanted weight to the portions that might lead to an unsatisfying decision. In some situations, architecture surfaces expose multiple materials with overlapping pathologies caused by numerous facts, the expert would simplify the case and make a general assertation. If each possible pathology is mapped out automatically by conducting a swift survey activity using an object detection model, the outcome will be suggestive and helpful.

The model performance for pathology detection is also hard to evaluate. Numerically, the trained model achieved mean Average Precision (mAP) at the Intersection over Union of 0.5 (mAP_0.5) is 0.13, and the average mAP at different IoU thresholds from 0.5 to 0.95 in steps of 0.05 (mAP_0.5:0.95) is 0.04, with precision of 0.21 and 0.22 recall. This is partially resulted by the box annotation approach for object detection. Pathological instances usually feature irregular shapes and ambiguous boundaries, the mAP based on IoU can hardly represent the model behaviour.

The model's behaviour is expected to be justified in the practical scene where the moving camera will register the surroundings in each timeframe. With the camera moving, the changing perspective will allow the model to perform detection several times.

6. CONCLUSION

This paper presented practical tests of classification and object detection of pathological issues in architectural preservation, using deep learning methods. With the expectation of understanding how much the methods can be applied to detect aimed pathology as 'stuff', which differs from the common 'thing' object, and applicability in further usage in 3D mapping.

The expectation of the deep learning methods is primarily limited and defined by the category design. This categorising system has to associate the capability and characteristics of the image feature extractor with the visual patterns that indicate semantic meanings. In this paper, the used category is based on the ICOMOS glossary for stone deterioration, maintaining the semantic hierarchy. The strategy of adopting semantic hierarchy is expected to improve the eventual reliability of the model in future application practices.

The paper has provided a detailed procedure for the application of deep learning computer vision methods to the architecture preservation field, the results are instructive to the general workflow. The evaluation results of the models reveal the problems with the identification of pathology using RGB image data collected without standards for perspective, ground sampling distance, light control etc. In a common situation, 'generic' images captured by non-experts or robots may be used, with a reduced possibility of capturing a dataset with predefined criteria. Therefore, the establishment of categories in the first place should be clear, concise, and consistent. Before the dataset preparation process, a clear manual should be set up, for naming the files, definition and disambiguation for each category, standard for annotation, proof checking of the training set. In the meantime, the training process should align with the wanted application scene. Multiple technical details should be taken care of during the data acquisition process, for example, the light and shading, the ground sampling distances of the camera, and the photoshoot perspective.

Critics of the trials to detect architectural pathology occur when it comes to generalising the trained model capability to different architecture cases, materials etc. Multiple pathological concepts are closely related to specific causes and the characteristics of the material. In the broader range of pathologies, biological colonisation, and chromatic alteration might appear in different ways with regard to architecture cases and materials. Pitting and alveolization, peeling and cracks share similarities in visual features. Therefore, it requires more data and further fitting of the model to examine the potential of the models. Considering additional information types, such as ultraviolet, infrared etc., with model complexity increased, the outcome will be improved. Another possible solution is to train a model for each specific case, with regards to materials and architectural cases.

The eventual goal of the test of object detection is to build a base for 3D mapping. The trained model for object detection acquired acceptable results. Biological colonisation and plant are the most distinct categories. Confusing pathologies like cracks and peelings can be well recognized. The characteristic of the pathology is that it cannot be easily defined by the boundary and is featured with irregular shapes. This characteristic is closely related to the definition of 'thing' and 'stuff' in the deep learning computer vision field. Bounded by this, the outcome of object detection with numerical value cannot fully represent the applicability of the methods. In this case, segmentation appears to be a better solution. However, considering that the corresponding decay areas are usually irregular, and the boundaries related to some pathologies are not well defined, the manual annotation for preparing the segmentation dataset will be costly.

ACKNOWLEDGEMENTS

The authors would like to thank especially Prof. Sonia Pistidda for her help in defining the semantic hierarchy of the category. Thanks would also be given to Chang He who for his help in data annotation. Thanks would also be to Prof. Cristian Campanella, Francesco Augelli, and Jiang Li for their generosity in sharing the data.

Financial support from the program of the China Scholarships Council (grant number: 202208520007) is acknowledged.

REFERENCES

Caesar, H., Uijlings, J., Ferrari, V., 2018. COCO-Stuff: Thing and Stuff Classes in Context. Proceedings of the IEEE conference on computer vision and pattern recognition, 1209-1218. doi.org/10.48550/arXiv.1612.03716

Charles, R.Q., Su, H., Kaichun, M., Guibas, L.J., 2017. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, 77–85. doi.org/10.1109/CVPR.2017.16

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, 248–255. doi.org/10.1109/CVPR.2009.5206848

Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A., 2010. The Pascal Visual Object Classes (VOC.) Challenge. Int. J. Comput. Vis. 88, 303–338. doi.org/10.1007/s11263-009-0275-4

Grilli, E., Dininno, D., Petrucci, G., Remondino, F., 2018. From 2D to 3D Supervised Segmentation and Classification for Cultural Heritage Applications. I*nt. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLII-2, 399–406. doi.org/10.5194/isprs-archives-XLII-2-399-2018

Guerrieri, M., Parla, G., 2022. Flexible and stone pavements distress detection and measurement by deep learning and low-cost detection devices. *Eng. Fail. Anal.* 141, 106714. doi.org/10.1016/j.engfailanal.2022.106714

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep Residual Learning for Image Recognition. doi.org/10.48550/arXiv.1512.03385

Kwon, D., Yu, J., 2019. Automatic Damage Detection of Stone Cultural Property Based on Deep Learning Algorithm. I*nt. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.*, XLII-2/W15, 639–643. doi.org/10.5194/isprs-archives-XLII-2-W15-639-2019

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., Jackel, L., 1989. Handwritten Digit Recognition with a Back-Propagation Network, in: NIPS.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P., 2015. Microsoft COCO: Common Objects in Context. doi.org/10.48550/arXiv.1405.0312

Maaten, L. van der, Hinton, G., 2008. Visualising Data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605.

Malinverni, E.S., Pierdicca, R., Paolanti, M., Martini, M., Morbidoni, C., Matrone, F., Lingua, A., 2019. Deep Learning for Semantic Segmentation of 3D Point Cloud. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.*, XLII-2/W15, 735–742. doi.org/10.5194/isprs-archives-XLII-2-W15-735-2019

Mishra, M., Barman, T., Ramana, GV, 2022. Artificial intelligence-based visual inspection system for structural health monitoring of cultural heritage. *J. Civ. Struct. Health Monit.*, 1-18. doi.org/10.1007/s13349-022-00643-8

Muhammad, M.B., Yeasin, M., 2020. Eigen-CAM: Class Activation Map using Principal Components. 2020 International Joint Conference on Neural Networks (IJCNN), 1–7. doi.org/10.1109/IJCNN48605.2020.9206626

Pathak, R., Saini, A., Wadhwa, A., Sharma, H., Sangwan, D., 2021. An object detection approach for detecting damages in heritage sites using 3-D point clouds and 2-D visual data. *J. Cult. Herit.*, 48, 74–82. doi.org/10.1016/j.culher.2021.01.002

Pierdicca, R., Paolanti, M., Matrone, F., Martini, M., Morbidoni, C., Malinverni, E.S., Frontoni, E., Lingua, A.M., 2020. Point Cloud Semantic Segmentation Using a Deep Learning Framework for Cultural Heritage. *Remote Sens.*, 12, 1005. doi.org/10.3390/rs12061005

Qi, C.R., Liu, W., Wu, C., Su, H., Guibas, L.J., 2018. Frustum PointNets for 3D Object Detection from RGB-D Data. Proceedings of the IEEE conference on computer vision and pattern recognition, 918-927. doi.org/10.48550/arXiv.1711.08488

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You Only Look Once: Unified, Real-Time Object Detection. Proceedings of the IEEE conference on computer vision and pattern recognition, 779-788. doi.org/10.48550/arXiv.1506.02640

Ren, S., He, K., Girshick, R., Sun, J., 2016. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. doi.org/10.48550/arXiv.1506.01497

Simonyan, K., Vedaldi, A., Zisserman, A., 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. doi.org/10.48550/arXiv.1312.6034

Simonyan, K., Zisserman, A., 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. doi.org/10.48550/arXiv.1409.1556

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2014. Going Deeper with Convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition, 1-9. doi.org/10.48550/arXiv.1409.4842

Teruggi, S., Grilli, E., Russo, M., Fassi, F., Remondino, F., 2020. A Hierarchical Machine Learning Approach for Multi-Level and Multi-Resolution 3D Point Cloud Classification. *Remote Sens.*, 12, 2598. doi.org/10.3390/rs12162598

Verges-Belmin, V., Stone (ISCS), ISC for, 2008. Illustrated glossary on stone deterioration patterns. ICOMOS.

Zhang, Kai, Teruggi, S., Ding, Y., Fassi, F., 2022. A Multilevel Multiresolution Machine Learning Classification Approach: A Generalization Test on Chinese Heritage Architecture. *Heritage*, 5, 3970–3992. doi.org/10.3390/heritage5040204

Zhang, K., Teruggi, S., Fassi, F., 2022. Machine Learning Methods for Unesco Chinese Heritage: Complexity and Comparisons. I*nt. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLVI-2/W1-2022, 543–550. doi.org/10.5194/isprs-archives-XLVI-2-W1-2022-543-2022

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning Deep Features for Discriminative Localization. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2921–2929. doi.org/10.1109/CVPR.2016.319