EXPLAINING AI: UNDERSTANDING DEEP LEARNING MODELS FOR HERITAGE POINT CLOUDS

EXPLAINING AI: UNDERSTANDING DEEP LEARNING MODELS FOR HERITAGE POINT CLOUDS

F. Matrone [1,*], A. Felicetti [2], M. Paolanti [3], R. Pierdicca [4]

[1] Department of Environmental, Land and Infrastructure Engineering (DIATI) - Politecnico di Torino, 10129, Torino, Italy – francesca.matrone@polito.it
[2] Department of Information Engineering (DII) - Università Politecnica delle Marche, 60131 Ancona, Italy - a.felicetti@staff.univpm.it
[3] Vision Robotics and Artificial Intelligence Lab (VRAI), Department of Political Sciences, Communication and International Relations, - University of Macerata, 62100 Macerata, Italy - marina.paolanti@unimc.it
[4] Department of Civil, Building Engineering (DICEA) - Università Politecnica delle Marche, 60131 Ancona, Italy - r.pierdicca@staff.univpm.it

KEY WORDS: Explainability, semantic segmentation, cultural heritage, point clouds

ABSTRACT:

Deep Learning has been pivotal in many real-world applications (e.g., autonomous driving, medicine and retail). With the wide availability of consumer-grade depth sensors, acquiring 3D data has become more affordable and effective, and many 3D datasets are currently publicly available. 3D data provides a great opportunity for a better comprehension of the surrounding environment for machines. There is a growing need for innovative methods for the treatment and analysis of point clouds and for their classification. The complex hidden layers, which are at the basis of deep neural networks (DNNs), make it difficult to interpret these models, that up to a few years ago DNNs were considered and treated as black box operators. Still, with their increasing popularity, making them explainable and interpretable has become mandatory. A lot of efforts were devoted to developing an Explainable Artificial Intelligence (XAI) framework for explaining DNNs decisions with 2D data, while only a few studies have attempted to investigate the explainability of 3D DNNs and, even more, heritage scenarios. To overcome these limitations, it was proposed the BubblEX framework: a novel multimodal fusion framework to learn the 3D point features. In our work, BubblEX has been exploited to understand the decisions taken by DNNs for heritage point clouds. The approach has been applied to a Digital Cultural Heritage Dataset, which is publicly available: the ArCH (Architectural Cultural Heritage) Dataset.

1. INTRODUCTION

Cultural Heritage (CH), both movable and immovable, namely built, architectural, natural, and landscape heritage, artworks and objects that express beauty or cultural values, seems to be the center of two contradictory nodal situations. On the one hand, thanks to the development of digital technologies and the effort of national and international projects, and local stakeholders involved in digitalization projects, we have faced a constant growth of digital material in the last decades (Yu et al., 2022). On the other hand, the current production workflow still limits access, generation, and use of such content (Argyrou and Agapiou, 2022).

Digital material is not always easy to access due to several factors, including the unavailability of adequate Information Technology (IT) equipment and the lack of citizens' interest and involvement (Bombini et al., 2022). The increasing availability of three-dimensional (3D) data, deriving from LiDAR (Light Detection And Ranging), MMSs (Mobile Mapping Systems) or UAVs (Unmanned Aerial Vehicles), provides the opportunity to rapidly generate detailed 3D scenes to support restoration, conservation, and safeguarding activities of built heritage. In the context of production, HBIM (Historic Building Information Modeling) constitutes a reference. Unlike the standard BIM (Building Information Modeling) methodology, where constructions are handcrafted by a designer, it applies a reverse engineering approach, typically relying on point clouds to perform the scan-to-BIM processes. These processes are still mostly manually carried out by domain experts (Pan and Zhang, 2022), making the workflow very time-consuming, not fully exploiting the potential of point clouds to automatically derive parametric objects, their segmentation, eventual physical issues, or any related metadata. Moreover, after constructing such models, their use in a real-time context is limited by the device's capabilities to render, transmit and analyze point clouds with potentially enormous sizes.

Deep neural network (DNN) models are currently adopted in several domains, from medical diagnosis (Fiorentino, 2022) to retail (Rossi et al., 2021) due to their ability to learn meaningful information from data and due to their success in many computer vision tasks (Xiao et al., 2018), where solid literature in the last years showed its potential. Deep Learning (DL) can enable the automatic recognition of architectural elements from point clouds. Various 2D vision problems have been successfully tackled with this dominating technique in AI. Semantic segmentation, shape, and surface detection are well studied in images and have been extended to the CH domain (Zhang et al., 2021), with 3D shape classification (Grilli and Remondino, 2019), 3D object detection and tracking (Fiorucci et al., 2020), and 3D point cloud segmentation (Malinverni et al., 2019). However, DNNs on heritage point clouds are still in their infancy due to the unique challenges given by the order-less nature of point clouds (Matrone et al., 2020a).

Initially, these models were considered and treated as black box operators, but with their increasing popularity, it becomes mandatory to make DNNs explainable and interpretable. This issue was considered a downside of deep learning for several years and users are generally reluctant to use techniques that are not fair and trustworthy, following the European movement to

pursue sustainable and ethical development of Artificial Intelligence (AI) (Goodman and Flaxman, 2017).

In this regard, effective and trustworthy DL algorithms are essential for Explainable AI (XAI). Explainability is a closely related concept to interpretability. Whereas interpretability focuses on abstract topics, explainability is the identification of pertinent features in the interpretable field that are important for accomplishing a peculiar decision (Tioa and Guan, 2020). Explanatory artificial intelligence tackles the critical issue that complex machines and algorithms often cannot provide insights into their behavior and thought processes. XAI allows users and parts of the internal system to be more transparent, providing explanations of their decisions in some level of detail. These explanations are crucial to ensure algorithmic fairness, identify potential biases or inconsistencies in the training data, and ensure that the algorithms perform as expected.

In recent years, much effort has been devoted to developing XAI methods for explaining DNN decisions, especially for 2D data (Graziani et al., 2020). XAI techniques are based on saliency maps, which denote the pixels deemed essential for the model's decision under consideration (Kindermans et al., 2019). Although there are many state-of-the-art studies on the exploitation of XAI techniques with 2D data (Young et al., 2019), few works have tackled examining the explainability of 3D DNNs (Zhao et al., 2020). In our previous work (Matrone et al., 2022), it has been proposed BubblEX, a multimodal fusion framework to learn the 3D point features, and it has been applied to Modelnet40 and ScanObjectNN datasets with suitable results. Its goal is to unfold the black box for the 3D point cloud features learning.

In this paper, we started from the BubblEX framework (Matrone et al., 2022) to explore the field of XAI for 3D heritage data. Among the various state-of-the-art DNNs, the Dynamic Graph CNN (DGCNN) (Wang et al., 2019) has been selected on the basis of previous works (Matrone et al., 2020a) and to ensure study continuity. Nevertheless, the proposed approach can be extended to any other architecture, being independent of the type of DNN chosen.

The DGCNN builds dynamic connections among points in their feature level and updates point features based on their neighboring points in the feature space. It has been trained on the ArCH dataset (Matrone et al., 2020b), in order to test the proposed methodology on heritage scenarios. In this context, the authors would like to investigate the potentialities of a domain shift of the XAI techniques in the cultural heritage sector, in addition to understanding why and how these approaches could be useful. Moreover, the differences between their application with other objects and categories are explored too.

The specific contributions of this paper are:

- the extension of the BubblEX framework to understand the process of 3D heritage point cloud features learning for multiclass scene understanding and interpreting;
- the implementation of a method developed for obtaining saliency maps from image data to deal with 3D heritage point cloud data;
- a visual method that enables analyzing and comparing multiple features;
- the generation of visual explanations from any DNN-based network for 3D heritage point cloud segmentation without requiring architectural changes or re-training.

The paper is organized as follows: section 1.1 provides a description of the state-of-the-art approaches developed for XAI on 3D point clouds. Section 2 describes the methodology according to the three different modules of BubblEX, namely the Learning phase (section 2.1), Visualization module (section 2.2), and Interpretability module (section 2.3). For better comprehension, the results of the Visualisation module have been included in the related section of the Methodology. In Section 3, Results and Discussions, an evaluation of our approach with respect to the ArCH dataset is offered, as well as a detailed analysis of the most relevant classes. Finally, in Section 4, final discussions on the obtained results are drawn, along with the conclusions and the definition of the future directions for this field of research.

1.1 Related works

In the literature, few studies have attempted to investigate the explainability of 3D DNNs. This section briefly reviews some relevant background works concerning XAI techniques for point cloud DNNs.

Zhang et al. proposed an explainable machine learning method called the PointHop (Zhang et al., 2020). It was specifically designed for point cloud classification task. PointHop built attributes of higher dimensions at each sampled point through iterative one-hop information exchange. This solution was like a larger receptive field in deeper convolutional layers in CNNs. The obstacle of unordered point cloud data was addressed by the authors with the adoption of a space partitioning procedure. Besides, it was applied the Saab transform to diminish the attribute dimension in each PointHop unit. In the classification phase, the feature vector was fed to a classifier, and the possibility of using ensemble methods to enhance the classification performance was explored.

In (Zheng et al., 2019), the authors have chosen to exploit the saliency map concept. In particular, they developed a saliency map for 3D point clouds to measure the importance of each point in a point cloud scene to model prediction loss. By approximating point dropping with a continuous point-shifting operation, they have shown that the contribution of a point was roughly proportional to, and thus can be scored by, the gradient of the loss with respect to the point under a scaled spherical-coordinate system. By using the saliency map, it was possible to standardize the point-dropping process to verify the veracity of the obtained saliency map on characterizing point-level and subset-level saliency.

However, the lead study on the use of XAI approaches to point clouds was (Gupta et al., 2020). It continues to be crucial for the comprehension of the feature sparsity of 3D models. The authors only reported sparse explanations that emphasize the influence of points at edges and corners, which is a lack of semantics, and the evaluation criterion of the explanations was not present. Besides, the gradient-based methods were not adjusted to models without gradients, such as tree-based models. Another evaluation approach was proposed by Adebayo et al. in (Adebayo et al., 2018). The authors randomized the network weights as well as the labels. Moreover, in their paper, they also stated that a feasible explanation should be sensitive to the weights of models and the data generating process.

In (Tan and Kotthaus, 2022), the authors proposed a solution for explaining the decision of a DNN when it deals with 3D data. In this paper, it has been described a point cloud-applicable XAI method based on a local surrogate model-based approach to determine which components are accountable for the classification. Furthermore, they quantified the efficacy of the explanations for point cloud data through fidelity and accuracy verification methods instead of a subjective approach based on human perception.

Considering the state-of-the-art in this context, the BubblEX framework comprises a visualization phase, followed by a recognition phase in which the important features for DNNs decisions are emphasized.

2. METHODOLOGY

The general structure of BubblEX for learning the features of 3D points is illustrated in Figure 1. Its structure consists of two modules, which follow the Learning phase: a) Visualization module and b) Interpretability module.

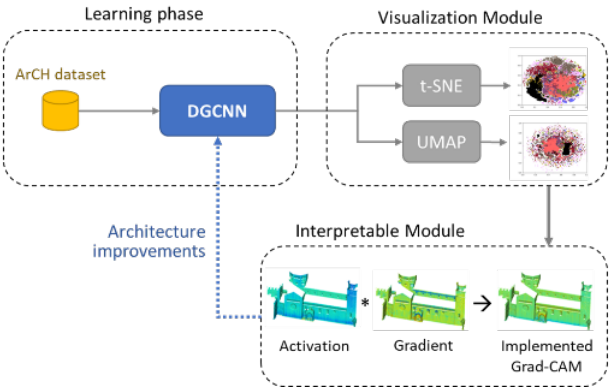

Figure 1. BubblEX workflow. The features extracted from the trained DNN serve as input for the Visualization and Interpretability Modules. The output of these two core parts could help to improve network decisions.

At first, a trained network on a point cloud dataset is required to solve a semantic segmentation task; then a method is selected for extracting and displaying the features learned from the network layers. At this stage, the insight resides on t-SNE (Van der Maaten and Hinton, 2008) and UMAP (McInnes et al., 2018), which are primarily designed to group neighbouring data points together. Aware of their criticalities (Chiari et al., 2021), in this case, they are mainly used to identify intruders among objects (Matrone et al. 2022) or clusters of points belonging to single classes (semantic segmentation task), demonstrating a good effectiveness for the direct visualization of the critical issues within the classes themselves. BubblEX adapts them to 3D point clouds. Finally, the third step is essential to understand the decisions made by the network to classify the features extracted in a given class. The union of these last two steps represents the fundamental fulcrum of the interpretability of the model trained in the initial steps. In fact, it allows to understand the decisional errors undertaken by the network and consequently could provide an idea of: a) how to improve the training phase, b) the accuracy of the dataset itself, and c) any critical issues due to the data acquisition phases. Finally, the Interpretability Module describes how adjacent points are involved in the extraction of features. For the development of this module, BubblEX was inspired by the Grad-CAM approach (Selvaraiu et al., 2017).

2.1 Learning phase

The DGCNN has been trained with different combinations of scenes belonging to the ArCH dataset, always excluding the test ones from the training and validation set. In this paper, the results obtained for a part of the *Trento square* scene (Figure 2) are shown. The network has been trained with the basic features (coordinates and RGB) in order to have a standard configuration similar to those of the state-of-the-art, and it performed with 84,62% of overall accuracy and 59,8% IoU.
The abovementioned network takes as input an n*3 tensor, which corresponds to the number of points in the batch (n) and the 3 coordinates. Each intermediate layer performs the feature extraction operations (n points*m features) implemented by the EdgeConv and Conv layers. Finally, the last layer for the classification (n*c), i.e. for each point, the layer assigns a vector of 9 elements corresponding to the likelihood of belonging to a class. During forward propagation, maps of activations are extracted, while during backward propagation, gradients are extracted. In particular, the gradients are computed in the last convolutional layer (*conv8*) before the output layer (Figure 3).

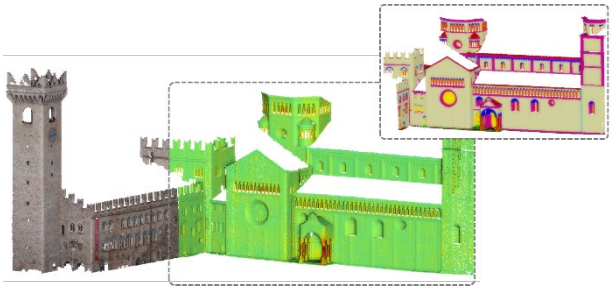

Figure 2. Trento square scene plotted and visualized by the RGB (left), implemented Grad-CAM (in the middle), and ground truth classes (top right).

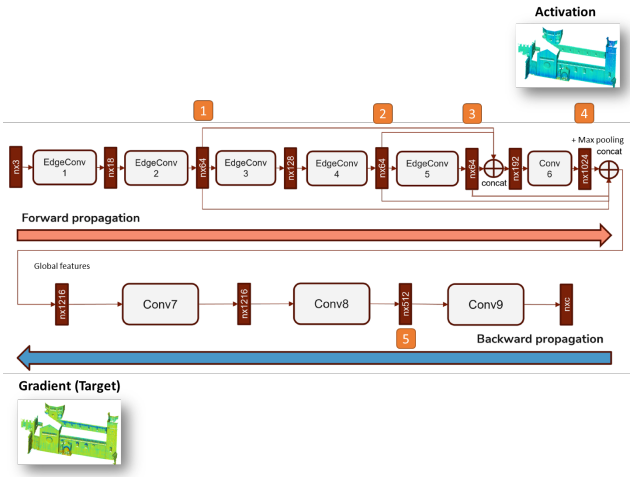

Figure 3. Numbers from 1 to 5 show the output layers for feature extraction. Point 1 and 5 are those selected for the Visualization Module, while only 5 for the Interpretation Module.

2.2 Visualization Module

To visualize the features learned by the network in its hidden layers, the large dimensionality of the data has to be taken into account. The state-of-the-art dimensionality reduction technique is the t-Distributed Stochastic Neighbor Embedding (t-SNE): particularly suitable for displaying features and large data sets. It is often used in the image domain, but in recent years it has also been successfully applied to other types of data, such as point clouds. With respect to PCA (Principal Component Analysis), which is a statistical technique, t-SNE is a probabilistic one. One of the main issues to be taken into account is its computational cost when dealing with high-dimensional data. The solution has been to apply the PCA as a dimensionality reduction technique, retaining at the same time most information. In recent years a new dimensionality reduction technique has been introduced, called Uniform Manifold Approximation and Projection (UMAP). UMAP is a learning technique for dimensional reduction based on Riemannian geometry and algebraic topology. This procedure is better than t-SNE in terms of reducing the dimensionality and display quality, as it allows for preserving most of the relations between the input data, providing fast processing times. These approaches allow understanding if the network is discriminating

well the different classes of the dataset within its architecture, in which way the network associates the wrong feature, and to investigate which are the closest points (consequently classes). Basically, each point in the dataset is associated with a feature vector extracted from an intermediate layer. This vector can be provided as input for one of these techniques, which will map it as a point in a 2D space. The whole test dataset will first be provided as input to the neural network, then transformed into feature vectors, and finally mapped within a two-dimensional space to be analyzed (Figure 4).

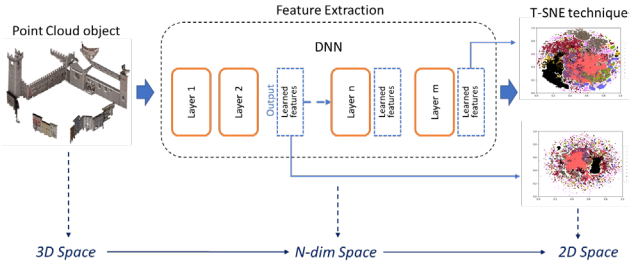

Figure 4. Workflow of spatial transformations of the data flow.

In detail, it starts with a 3D object which is given as a point cloud, with points represented as coordinates or other features. Within every single layer, new features are learned and defined, able to describe the single point with respect to its neighbors or the class/object in its entirety. However, the extraction of the features from different layers could be difficult to investigate, due to the n-dimensionality, therefore, they are plotted into a 2D space. For the extraction of the learned features, the eighth convolutional layer was chosen since the ninth is the output one. It has to be noticed that, in the case of point cloud classification, each point plotted in the t-SNE or UMAP corresponded to an object of the dataset (chair, monitor, wardrobe, plant, curtain, etc.) (Figure 5); while in the case of semantic segmentation, this kind of visualization was not possible. In fact, each class contains multiple objects, even separated from each other in real space, and it has been thus not feasible to plot the individual objects, within the same category, with a single point. Therefore, differently from what is shown in (Matrone at al., 2022), the points of the graphs actually represent all the points of the analyzed test scenes. However, this does not affect the overall effectiveness of the proposed method since it is still possible to identify, if the training phase is correct, the clusters of the single classes.

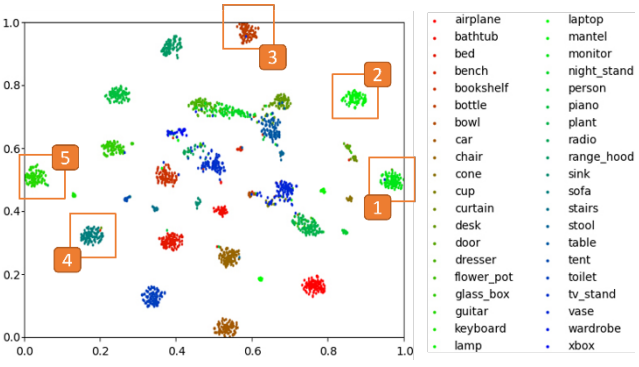

Figure 5. Results of t-SNE on the ModelNet40 dataset (classification task). Clusters have been clearly isolated, and within some of them (orange rectangle), "intruder" objects are visible, namely objects with features similar to those of the cluster, but predicted as another class (Matrone et al., 2022).

The results of the feature visualization using the t-SNE and UMAP techniques are depicted in Figure 6 and 7, where the point color represents the class and could be plotted according to the comparison with the ground truth (GT) or the predicted class. The proximity of the points in a single cluster or area of the graph indicates that the features in the feature space (learned from the network) are similar and, therefore, able to discriminate the object or class properly. The presence of points with different colors within a cluster demonstrates how the object (in the case of classification) or the point itself (in the case of segmentation) has been incorrectly predicted since the features describing it are similar to those of other objects. At this point, it is possible to investigate the cause in two ways: a) with the second module of the framework, visualizing directly on the point cloud the parts that have been mainly considered by the network; b) by directly examining the object or class to verify the absence of errors in the initial labeling phases of the dataset or that potential similarity of geometries to those of other classes.

If we compare Figure 5 with 6 and 7, we immediately see how in the case of the ArCH dataset, the final result is much more uneven and with clearly less defined clusters than those obtained with ModelNet40. The reasons are essentially two. First, the plotted points indicate individual points in the scene and not objects in the dataset, so there are more items plotted. Secondly, the network performances obtained with the ArCH dataset were lower than those obtained with ModelNet40. This last element is very relevant, as the difficulties and inhomogeneities of cultural heritage datasets are well known and their correct semantic classification is not straightforward and yet fully solved. In fact, the t-SNE and UMAP graphs of other 3D datasets, also addressed to the semantic segmentation task, gave clearer results with mainly separated clusters.

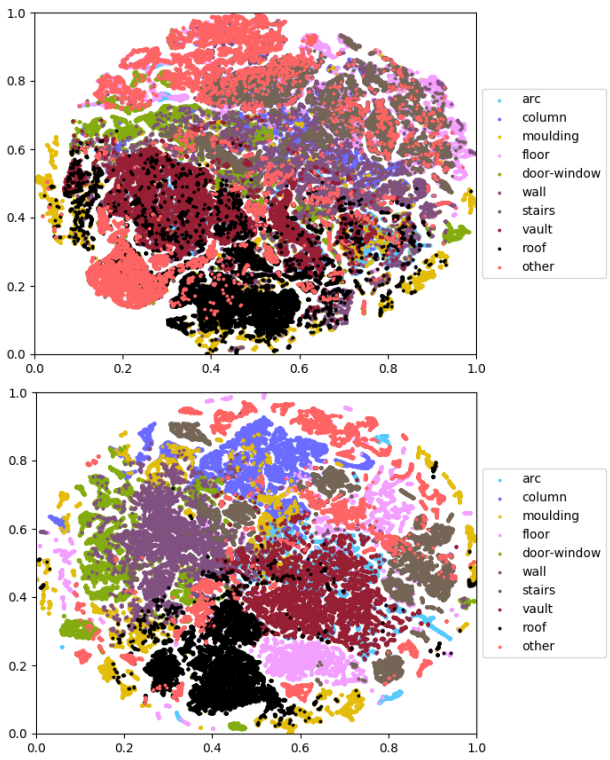

Figure 6. t-SNE of the test scene of the ArCH dataset plotted with respect to the prediction. At the top, the features extracted from the first convolutional layer, and at the bottom from the last before the final classification.

In detail, from Figure 6 it can be noticed that in the initial layer, there are many points predicted as "other", which are considerably reduced in the final layer, in which they are classified as different categories. This step indicates a good ability of the network to learn and discriminate, although not yet optimal. It is therefore not possible, in this case, to proceed as with ModelNet40, identifying the individual misclassified objects or points; however, it is clear that some classes such as "wall", "column", and "floor" are gradually described with more similar features. It can thus be assumed that the performance of the network can improve with a parallel increase of the dataset.

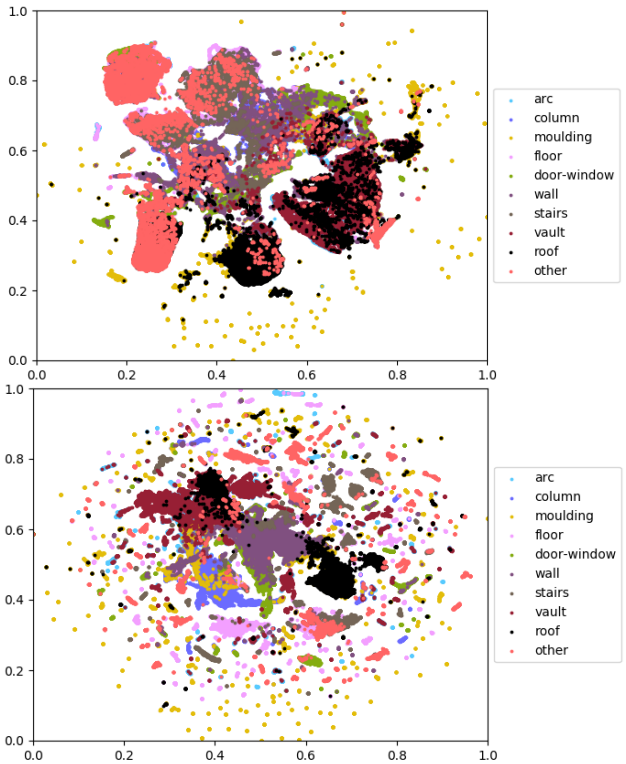

Figure 7. UMAP of the test scene of the ArCH dataset plotted with respect to the prediction. At the top, the features extracted from the first convolutional layer, and at the bottom from the last before the final classification.

Different results have been obtained for Figure 7, in which split and randomly distributed points are clearly visible. This behavior is intrinsically due to the technique itself, which tends to separate clusters more than t-SNE. From here, it is evident how, above all, the "molding" class is not described with homogeneous features, as well as the "arch", returning, in fact, the worst metrics.

2.3 Interpretability module

In this module, the activations in the innermost layers are analyzed along with Grad-CAM. In particular, we analyze the activations and the Grad-CAM at the output of the penultimate convolutional layer *conv8* before the final classifier *conv9*. This layer has a dimension of 256 features * 4096 points.
The Grad-CAM, unlike the activation, which corresponds to the output of the layer as a function of the input, is parametric and must be calculated according to the target class. It is equivalent to the product between the activation and the gradient obtained from the back-propagation of the output (both of size 256 features * 4096 points). To compute the gradient, a one-hot-encoding signal identifying the target class is multiplied by the output vector and back-propagated to the *conv5* layer in analogy to the error back-propagation during network training.
This activation, namely a matrix of dimension 1024 features * 1024 points, is then multiplied with the gradients, obtained after the back propagation, to get the implemented Grad-CAM results (Figure 8).
The implementation of Grad-CAM for the 3D data as point clouds required the analysis of the best function to flatten the feature size, previously extracted, and the exploration of the best combination of activation and gradient. With respect to the average, the median applied to the gradient has a lower dispersion of the points around their central values (Matrone et al., 2022). This feature was then multiplied with activation to modulate its values (as represented in Figure 8).

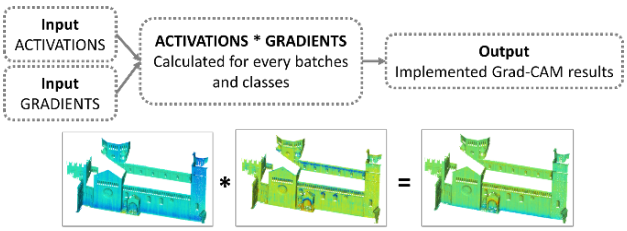

Figure 8. Interpretability module workflow.

Furthermore, since the use of the median to flatten both activation and gradient before multiplication did not conduct to a proper and immediate visualization, the median has been applied subsequently, succeeding in determining a significant variation compared to those of activation alone.
To facilitate the visualization, a colormap jet is used to emphasize the intensity of the values, where blue and red display the influential points, although their contribution is opposite. In fact, they map -1 and 1 respectively, while green is close to 0. The activations and gradients were calculated by iterating over the individual batch and then merging them together in the complete scene. Thanks to the activations taken in the forward propagation and the gradients extracted from the backward propagation, the Grad-CAM for entire scenes (Figure 9) for each class was calculated.
We can see how the product between the activations and the gradients highlights in the Grad-CAM the most considered parts of each class with the color red (+1) and the least significant parts for the classification of the same class with the color blue (-1) while non-influential points are represented with a color close to green (0).

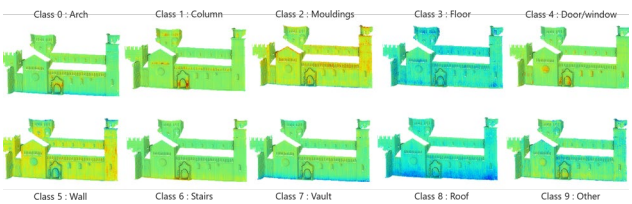

Figure 9. Implemented Grad-CAM for every class.

3. RESULTS AND DISCUSSIONS

The results of the implemented Grad-CAM made it possible to analyze the individual classes in detail and understand which elements the neural network considers to classify the points.

3.1 Columns

Figure 10 clearly depicts how the network locates the most relevant element in the shaft of the columns to distinguish this

class, partially omitting the capitals and the base. This behavior occurs both for columns with larger dimensions and with smaller ones, such as those of the colonnade of the upper loggia. In this case, the target class is 1, i.e. the columns, and both the GT class and the predicted one correspond to 1.

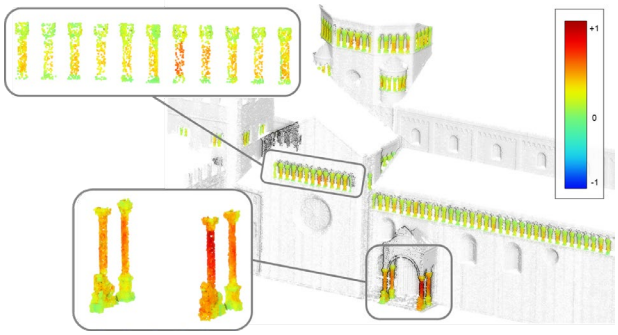

Figure 10. Column class with shafts highlighted. Target class 1 "Column" and GT class 1.

However, if we analyze another part of the façade, we can also notice how some points belonging to the column class have not been correctly predicted (Figure 11a). In this case, it has been useful to see their assigned class, and it resulted in class 5 "wall" (Figure 11b with target class 5 and predicted class 5). From a comparison with the RGB point cloud, it can be seen that these columns have not been completely reconstructed since the acquisition took place via terrestrial photogrammetry. Precisely, due to the partial absence of part of the geometry, the DNN could have been misled and, therefore, classified these columns as a wall. It is not thus a problem of learning and training, but of the incompleteness of the dataset, demonstrating how, for the column class, the partial absence of the curved element leads to misclassification.

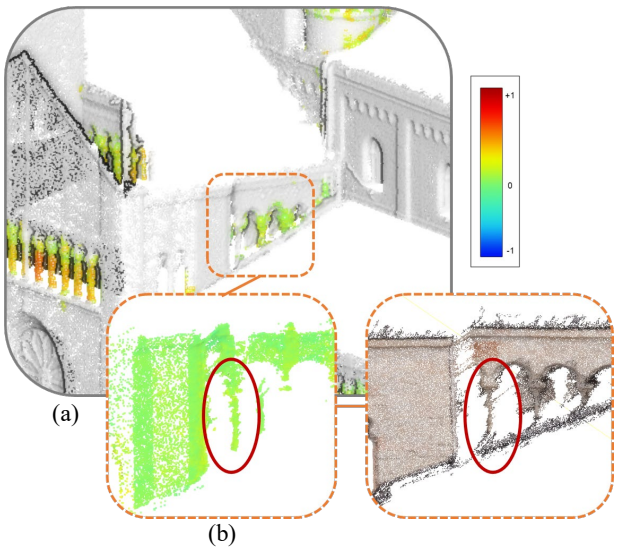

Figure 11 a and b. Column class with misclassified points. Target class 5 "Wall" and Predicted class 5.

3.2 Door and Windows

Also for this category, as for the previous one, the points correctly predicted with respect to GT were first considered, namely Target class 4 "Door-window" and GT class 4.
Figure 12a depicts how the central part of the windows, and openings in general, is mainly considered by the network. The red vertical bands, visible in the rosette, are almost certainly due to the ways in which the network analyses the scene: through vertical blocks that may have influenced the results. Despite this, it is clear that the outermost parts, sometimes in masonry, are secondary to the attention dedicated to the flat part of the openings. In fact, they have also been misclassified if having a look at Target class 4 - Predicted class 4 (Figure 12b). This result could lead to state that, for this class, acquisition via photogrammetry (both terrestrial and aerial) could be preferable to the laser scanning technique, where stained glass windows are rarely detected.

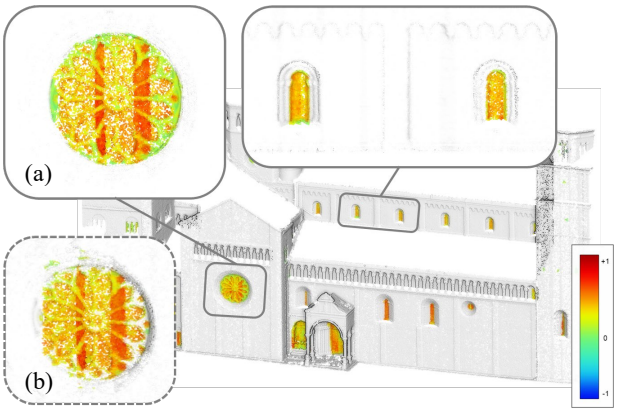

Figure 12. Door-Window class. (a) Target class 4 "Door-window" and GT class 4, (b) Target class 4 - Predicted class 4.

Figure 13 represents a comparison between Target class 4 - GT class 4 (Figure 13a) and Target class 4 - Predicted class 4 (Figure 13b). If the central windows indicated by the red circle are examined, it is possible to see that they are not recognized. In the first analysis, it was thought that this outcome was due to: a) the previously described cause, i.e. the method of analysis of the network using vertical blocks, whereby the points of the window had been separated into two different blocks and, consequently, not recognized or b) the fact that the windows resided approximately on the same plane as the masonry, less recessed than the others, and therefore recognized as a wall. However, later, the point clouds and the acquired images were reanalyzed and it was realized that these windows were the only ones walled up and closed (Figure 13c), therefore, the error was not from the DNN but from an incorrect dataset annotation.
Nevertheless, at this point, an open question arises: how should those two portions of the point cloud be annotated? Option 1: as *windows*, because they were openings and their shape is similar to the others, having only been walled-up? Or option 2: as *wall,* even if from a structural point of view they do not have the same behaviour as the adjacent masonry, thus following the result of the DNN?

Figure 13. Door-Window class with highlighted the misclassified openings and a comparison with the real state.

3.3 Arch and Stairs

Finally, the "arch" and "stair" classes have been analyzed. For the first, the central and most curved parts of these architectural elements have been highlighted by the implemented Grad-CAM (Figure 14), even if the DNN is sometimes confusing most external parts with the "wall" class. This result is justified by the low presence of arches in the ArCH dataset, among the classes with the least number of points.

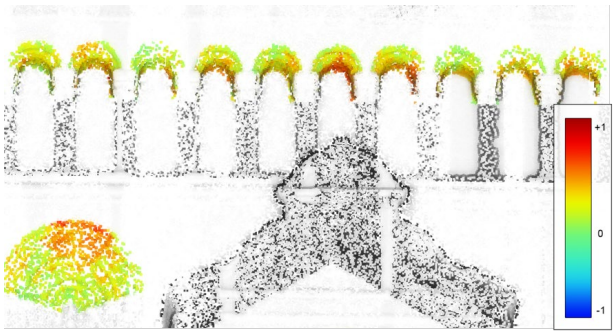

Figure 14. Target class 0 – GT class 0 *(Arch)*

For the second, on the other hand, equal importance was highlighted for the riser of the step with respect to the tread (Figure 15). This result could indicate how the geometric trend of the staircase is precisely the discriminating factor of this category. In fact, if only the riser were considered, it could be confused with a wall, if instead only the tread was considered, it could be misclassified with a floor, which, however, did not happen.

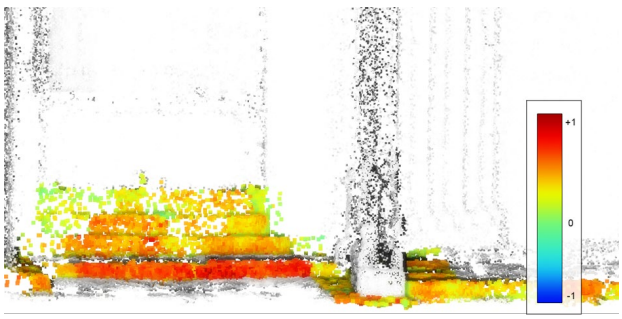

Figure 15. Target class 6 – GT class 6 *(Stairs)*

4. CONCLUSIONS

In conclusion, the application of some explainable approaches in the field of cultural heritage can, on the one hand, share common points with other domains (such as robotics, retail or medicine), but on the other, provide new insights domain specific. In particular, they may be useful to understand and evaluate the effectiveness of the code or the kind of features learned and their correctness, as well as for other sectors, but they can, above all, allow us to understand how to effectively acquire the data.

There is a plethora of tools and possibilities for surveying techniques, which today are integrated and combined to try to produce ever more complete data; however, there are situations in which the speed of the acquisition phases plays a crucial role. Examples are areas at risk (seismic, environmental or due to wars), where accurate survey campaigns cannot be carried out. In this case, knowing which are the fundamental parts or elements to be acquired for future automatic recognition of these components in augmented or virtual reality environments, as well as for digital reconstruction, could be of considerable help. Besides, the domain shift between everyday objects, such as those present in ModelNet40, to the point clouds of the architectural heritage could allow explainability techniques, in the next future, to fully and properly define the distinctive features of a heritage building, an architectural style or an architect, understanding their influences. In fact, their only application to images could be reductive, as architecture is an all-around work, unlike, for example, a painting (Diaz et al., 2020). In the present contribution, the saliency maps for some classes have been shown, demonstrating what a DNN focuses on; however, it is possible to imagine that with more extensive data, this methodology could be precisely applied to study architectural styles or historical influences, determining thus new interconnections and knowledge.

Furthermore, the scalability of these techniques has also been highlighted. In fact, to date, it has been applied to "simple" objects, e.g. monitors, curtains or vehicles, which usually maintain approximately the same dimensions. The case studies of cultural heritage, on the other hand, can greatly differ in size and geometry, even within the same category. If we consider the results obtained, it has been shown that the proposed methodology correctly works both on, for example, columns and openings with different dimensions but also shapes.

The main difference that is still present today in the application of explainability techniques to ordinary environments or objects with respect to the cultural heritage domain is the limited presence of labeled heritage point cloud datasets. The heterogeneity of the point clouds of heritage scenes and the scarcity of available data do not yet guarantee both the performances achieved in the other domains and the full development of these techniques.

Future development of this research will be the application of these approaches to heritage point clouds acquired with specific techniques, such as laser scanning, mobile mapping systems, and terrestrial or aerial photogrammetry, to understand if the predictions of the DDN would change according to the type of acquisition.

ACKNOWLEDGMENTS

The authors would like to acknowledge Lisa Burini, Matteo Ferretti, and Samuele Leli who supported part of the study.

This study was carried out within the FAIR - Future Artificial Intelligence Research and received funding from the European Union Next-GenerationEU (PNRR) – Missione 4 componente 2, investimento 1.3 – D.D. 1555 11/10/2022, PE00000013). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

REFERENCES

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M. and Kim, B., 2018. Sanity checks for saliency maps. *Advances in neural information processing systems,* 31.

Argyrou, A., Agapiou, A., 2022. A Review of Artificial Intelligence and Remote Sensing for Archaeological Research. *Remote Sensing*, 14(23), p.6000.

Bombini, A., Anderlini, L., dell'Agnello, L., Giaocmini, F., Ruberto, C. and Taccetti, F., 2022, May. The AIRES-CH Project: Artificial Intelligence for Digital REStoration of Cultural Heritages Using Nuclear Imaging and Multidimensional Adversarial Neural Networks. In *ICIAP 2022,* Lecce, Italy, May 23–27, 2022, *Proceedings, Part I,* pp. 685-700. Springer International Publishing.

Chari, T., Banerjee, J. and Pachter, L., 2021. The specious art of single-cell genomics. BioRxiv, pp.2021-08. doi.org/10.1101/2021.08.25.457696.

Díaz-Rodríguez, N. and Pisoni, G., 2020, July. Accessible cultural heritage through explainable artificial intelligence. In *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, pp. 317-324.

Fiorentino, M.C., Villani, F.P., Di Cosmo, M., Frontoni, E. and Moccia, S., 2022. A review on deep-learning algorithms for fetal ultrasound-image analysis. *Medical Image Analysis*, p.102629.

Fiorucci, M., Khoroshiltseva, M., Pontil, M., Traviglia, A., Del Bue, A. and James, S., 2020. Machine learning for cultural heritage: A survey. *Pattern Recognition Letters*, 133, pp.102-108.

Goodman, B. and Flaxman, S., 2017. European Union regulations on algorithmic decision-making and a "right to explanation". A*I magazine*, 38(3), pp.50-57.

Graziani, M., Andrearczyk, V., Marchand-Maillet, S. and Müller, H., 2020. Concept attribution: Explaining CNN decisions to physicians. *Computers in biology and medicine*, 123, p.103865.

Grilli, E. and Remondino, F., 2019. Classification of 3D digital heritage. *Remote Sensing*, 11(7), p.847.

Kindermans, P.J., Hooker, S., Adebayo, J., Alber, M., Schütt, K.T., Dähne, S., Erhan, D. and Kim, B., 2019. The (un) reliability of saliency methods. *Explainable AI: Interpreting, explaining and visualizing deep Learning*, pp.267-280.

Gupta, A., Watson, S. and Yin, H., 2020, July. 3D point cloud feature explanations using gradient-based methods. In *2020 International Joint Conference on Neural Networks* (IJCNN), pp. 1-8.

Malinverni, E.S., Pierdicca, R., Paolanti, M., Martini, M., Morbidoni, C., Matrone, F. and Lingua, A., 2019. Deep Learning for semantic segmentation of 3D point cloud. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences,* Volume XLII-2/W15, 2019 27th CIPA International Symposium "Documenting the past for a better future", 1–5 September 2019, Ávila, Spain.

Matrone, F., Grilli, E., Martini, M., Paolanti, M., Pierdicca, R. and Remondino, F., 2020a. Comparing machine and deep learning methods for large 3D heritage semantic segmentation. *ISPRS International Journal of Geo-Information*, 9(9), p.535. https://doi.org/10.3390/ijgi9090535

Matrone, F., Lingua, A., Pierdicca, R., Malinverni, E.S., Paolanti, M., Grilli, E., Remondino, F., Murtiyoso, A. and Landes, T., 2020b. A benchmark for large-scale heritage point cloud semantic segmentation. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43, pp.1419-1426.

Matrone, F., Paolanti, M., Felicetti, A., Martini, M. and Pierdicca, R., 2022. BubblEX: An Explainable Deep Learning Framework for Point-Cloud Classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15, pp.6571-6587.

McInnes, L., Healy, J. and Melville, J., 2018. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

Pan, Y. and Zhang, L., 2022. Integrating BIM and AI for Smart Construction Management: Current Status and Future Directions. *Archives of Computational Methods in Engineering*, pp.1-30.

Paolanti, M. and Frontoni, E., 2020. Multidisciplinary pattern recognition applications: A review. *Computer Science Review*, 37, p.100276.

Rossi, L., Paolanti, M., Pierdicca, R. and Frontoni, E., 2021. Human trajectory prediction and generation using LSTM models and GANs. *Pattern Recognition*, 120, p.108136.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).

Tan, H. and Kotthaus, H., 2022. Surrogate model-based explainability methods for point cloud NNs. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision,* pp. 2239-2248.

Tjoa, E. and Guan, C., 2020. A survey on explainable artificial intelligence (XAI): Toward medical xai. IEEE transactions on neural networks and learning systems, 32(11), pp.4793-4813.

Van der Maaten, L. and Hinton, G., 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Xiao, Y., Wu, J., Lin, Z. and Zhao, X., 2018. A deep learning-based multi-model ensemble method for cancer prediction. *Computer methods and programs in biomedicine*, 153, pp.1-9.

Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M. and Solomon, J.M., 2019. Dynamic graph CNN for learning on point clouds. *Acm Transactions On Graphics*, 38(5), pp.1-12.

Young, K., Booth, G., Simpson, B., Dutton, R. and Shrapnel, S., 2019. Deep neural network or dermatologist?. In I*nterpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support: Second International Workshop, iMIMIC 2019, and 9th International Workshop, ML-CDS 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Proceedings 9* (pp. 48-55). Springer International Publishing.

Yu, T., Lin, C., Zhang, S., Wang, C., Ding, X., An, H., Liu, X., Qu, T., Wan, L., You, S. and Wu, J., 2022. Artificial Intelligence for Dunhuang Cultural Heritage Protection: The Project and the Dataset. *International Journal of Computer Vision*, 130(11), pp.2646-2673.

Zhang, M., You, H., Kadam, P., Liu, S. and Kuo, C.C.J., 2020. Pointhop: An explainable machine learning method for point cloud classification. *IEEE Transactions on Multimedia*, 22(7), pp.1744-1755.

Zhang, R., Li, G., Wunderlich, T. and Wang, L., 2021. A survey on deep learning-based precise boundary recovery of semantic segmentation for images and point clouds. *International Journal of Applied Earth Observation and Geoinformation*, 102, p.102411.

Zhao, B., Hua, X., Yu, K., Tao, W., He, X., Feng, S. and Tian, P., 2020. Evaluation of Convolution Operation Based on the Interpretation of Deep Learning on 3-D Point Cloud. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13, pp.5088-5101.

Zheng, T., Chen, C., Yuan, J., Li, B. and Ren, K., 2019. Pointcloud saliency maps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision,* pp. 1598-1606.