Transfer learning and performance enhancement techniques for deep semantic segmentation of built heritage point clouds

Virtual Archaeology Review, **12**(25): 73-84, 2021
https://doi.org/10.4995/var.2021.15318

Received: March 24, 2021
Accepted: April 14, 2021

TRANSFER LEARNING AND PERFORMANCE ENHANCEMENT TECHNIQUES FOR DEEP SEMANTIC SEGMENTATION OF BUILT HERITAGE POINT CLOUDS

TRANSFERENCIA DE TÉCNICAS DE APRENDIZAJE Y MEJORA DEL RENDIMIENTO EN LA SEGMENTACIÓN SEMÁNTICA PROFUNDA DE NUBES DE PUNTOS DEL PATRIMONIO CONSTRUIDO

Francesca Matrone[a,*], Massimo Martini[b]

[a] Department of Land and Infrastructure Engineering (DIATI), Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy.
francesca.matrone@polito.it

[b] Department of Information Engineering (DII), Università Politecnica delle Marche, Via Brecce Bianche 12, 60100 Ancona, Italy.
m.martini@pm.univpm.it

Highlights:

- Semantic segmentation of built heritage point clouds through deep neural networks can provide performances comparable to those of more consolidated state-of-the-art ML classifiers.
- Transfer learning approaches, as fine-tuning, can considerably reduce computational time also for CH domain-specific datasets, as well as improve metrics for some challenging categories (i.e. windows or mouldings).
- Data augmentation techniques do not significantly improve overall performances.

Abstract:

The growing availability of three-dimensional (3D) data, such as point clouds, coming from Light Detection and Ranging (LiDAR), Mobile Mapping Systems (MMSs) or Unmanned Aerial Vehicles (UAVs), provides the opportunity to rapidly generate 3D models to support the restoration, conservation, and safeguarding activities of cultural heritage (CH). The so-called scan-to-BIM process can, in fact, benefit from such data, and they can themselves be a source for further analyses or activities on the archaeological and built heritage. There are several ways to exploit this type of data, such as Historic Building Information Modelling (HBIM), mesh creation, rasterisation, classification, and semantic segmentation. The latter, referring to point clouds, is a trending topic not only in the CH domain but also in other fields like autonomous navigation, medicine or retail. Precisely in these sectors, the task of semantic segmentation has been mainly exploited and developed with artificial intelligence techniques. In particular, machine learning (ML) algorithms, and their deep learning (DL) subset, are increasingly applied and have established a solid state-of-the-art in the last half-decade. However, applications of DL techniques on heritage point clouds are still scarce; therefore, we propose to tackle this framework within the built heritage field. Starting from some previous tests with the Dynamic Graph Convolutional Neural Network (DGCNN), in this research close attention is paid to: i) the investigation of fine-tuned models, used as a transfer learning technique, ii) the combination of external classifiers, such as Random Forest (RF), with the artificial neural network, and iii) data augmentation results evaluation for the domain-specific ArCH dataset. Finally, after analysing the main advantages and critical aspects, a proposal is made evaluating the extent to which this methodology can also be useful for non-programming or domain experts.

Keywords: cultural heritage; semantic segmentation; deep learning; deep neural networks; point clouds

Resumen:

La creciente disponibilidad de datos tridimensionales (3D), como nubes de puntos, provenientes de la detección de la luz y distancia (LiDAR), sistemas de mapeado móvil (MMS) o vehículos aéreos no tripulados (UAV), brinda la oportunidad de generar rápidamente modelos 3D para apoyar las actividades de restauración, conservación y salvaguardia del patrimonio cultural (CH). El llamado proceso de escaneado-a-BIM puede, de hecho, beneficiarse de dichos datos, y ellos mismos pueden ser una fuente para futuros análisis o actividades sobre el patrimonio arqueológico y el construido. Hay varias formas de explotar este tipo de datos, como el modelado de información de edificios históricos (HBIM), la creación de mallas, la rasterización, la clasificación y la segmentación semántica. Este último, referido a las nubes de puntos, es un tema de máxima actualidad no solo en el dominio del PC sino también en otros campos como la navegación autónoma, la medicina o el comercio minorista. Precisamente en estos sectores, la tarea de la segmentación semántica se ha explotado y desarrollado principalmente con técnicas de inteligencia artificial. En particular, los algoritmos de aprendizaje automático (AA) y su subconjunto de aprendizaje profundo (AP) se aplican cada vez más y han establecido un sólido estado de la técnica en la última media década. Sin embargo, las aplicaciones de las técnicas de AP en las nubes de puntos tradicionales son todavía escasas; por tanto, nos proponemos abordar este

*Corresponding author: Francesca Matrone, francesca.matrone@polito.it

marco dentro del ámbito del patrimonio construido. Partiendo de algunas pruebas anteriores con la Red Neural Convolucional de Gráfico Dinámico (DGCNN), en esta contribución se presta atención a: i) la investigación de modelos afinados, utilizados como técnica de aprendizaje por transferencia, ii) la combinación de clasificadores externos, como Random Forest (RF), con la red neuronal artificial, y iii) la evaluación de los resultados de aumentación de datos para el conjunto de datos específico del dominio ArCH. Finalmente, después de analizar las principales ventajas y los aspectos criticables, se hace una propuesta valorando hasta qué punto esta metodología puede ser útil también en expertos no programadores o del campo.

Palabras clave: patrimonio cultural; segmentación semántica; aprendizaje profundo; redes neuronales profundas; nubes de puntos

1. Introduction

In the Cultural Heritage (CH) field, point clouds are an increasingly used tool for asset management. The development in recent years of faster and more efficient acquisition tools such as Mobile Mapping Systems (MMSs) has contributed to the widespread use of these 3D data in several sectors such as autonomous navigation, robotics and augmented and virtual reality. In the Digital Cultural Heritage (DCH) domain, combining these systems with more consolidated techniques such as terrestrial laser scanners, terrestrial and aerial photogrammetry using UAVs (Unmanned Aerial Vehicle), allows the acquisition of massive amounts of data, sometimes even excessive. In fact, for their effective use, point clouds are usually subsampled, filtered and post-processed, in order to simplify their management.

In addition to these operations, a new trend has recently emerged: the semantic segmentation of point clouds through artificial intelligence techniques such as Machine and Deep learning (ML/DL). This tendency allows point clouds to be used as a basis for 3D modelling or as a support for semantic data processing. The subdivision of the point clouds into certain classes (for an architectural, archaeological, urban or regional scale) entails various tasks: speeding up the reconstruction of 3D models, such as BIM (Building Information Modelling) models (Bassier et al., 2020); automating analysis in Geographic Information Systems (GIS) environments, supporting 3D city modelling (Park and Gulmann, 2019); facilitating and integrating the representation of forms of decay (Grilli & Remondino, 2019) and so on. To foster research in this direction, it is necessary to implement an automatic semantic segmentation, even if the unstructuredness of point clouds makes use of DL not straightforward. In the computer vision, this task is now consolidated and well established in the literature for both 2D and 3D data. However, for the 3D data of architectural heritage, there is not yet a strong background.

With this research, we, therefore, aim to propose a new methodology for the semantic segmentation of heritage point clouds through DL techniques. In this way, it is possible to automate the recognition of the various architectural classes and overcome some limitations given by the use of 2D images such as incomplete data (given by the lack of three-dimensionality), lighting problems or possible occlusions. Besides, an attempt is made to increase the level of detail (LoD) achieved to date in the state-of-the-art for the semantic segmentation of point clouds (Weinmann et al., 2015; Boulch et al., 2018; Landrieu & Simonovsky, 2018). Among the usual and general classes as Building, Vegetation, Street or Vehicle, we would detail the Building class with Roof, Column, Moulding, Stair, Wall, Arch, Floor, Vault and Door/Window, for a total of nine subclasses. Finally, given some breakthroughs in the field of DCH with classifiers such as Random Forest (RF) (Teruggi et al., 2020), we propose a further comparison w.r.t. (Matrone et al., 2020a) between the DL and ML methodologies in order to complete the analysis framework, as well as study a method for their integration.

Within this contribution, which is part of the broader debate on Digital Humanities, three research questions are addressed:

- Is it possible to use DL techniques for the CH domain where the standardisation of the elements, which should help automatic recognition, is almost absent, thus making the task even more challenging?
- What are the pros and cons of the deep neural networks (DNNs) compared to the most consolidated ML classifiers?
- Is it possible to make the proposed methodology “user-friendly” for those who are not programming or domain experts?

2. State of the art

Since DL is a subset of ML, it is useful to examine how the overall framework is dealing with both 2D and 3D data of DCH, to subsequently detail only the DL.

2.1. The datasets

As stated by (Fiorucci et al., 2020) the application of ML to the field of CH is not yet fully widespread, and it is severely bounded by the lack of adequate datasets. Besides limiting the development of specific algorithms for DCH, this lack also prevents a full comparison of the different solutions proposed by the researchers. This absence of datasets drives the studies to mostly train DNNs on external datasets. Then, through a transfer learning approach, they use the last layers of the pre-trained network to save the features and implement a final fine-tuning based on a new smaller dataset, targeted on the case study under examination.

The issue of the dataset has a key role in determining the success of the DL framework for the CH domain. At the very beginning of this research, it was not possible to identify one suitable dataset for our purposes, hence it was necessary to create an *ad hoc* one. In fact, if in the case of 2D data there were (Korc and Förstner, 2009; Teboul et al., 2012; Tyleček and Šára, 2013), specific for some CH areas, but still inherent to the topic, for the 3D data the availability was limited to an urban scale or highly-serialised indoor environments such as offices. Examples of this datasets are Semantic3D (Hackel et al., 2017), S3DIS (Armeni et al., 2016) or KITTI (Geiger et al., 2013).

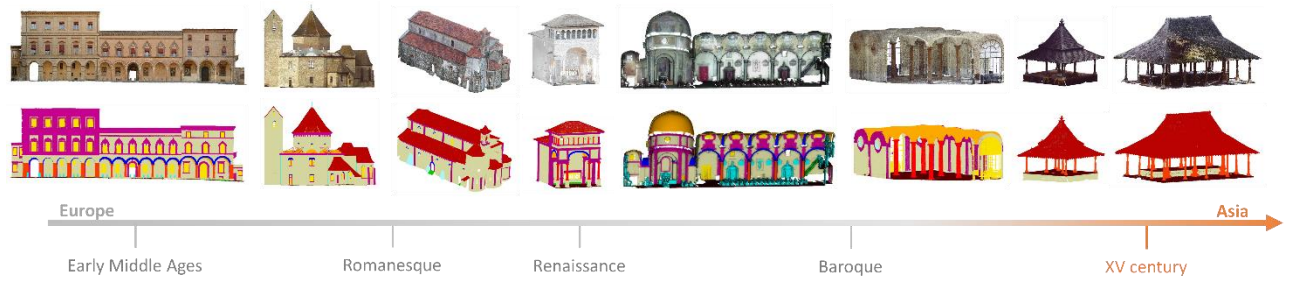


Figure 1: Different architectural styles and epochs of the CH point clouds within the ArCH dataset.

A new dataset has been therefore proposed, named ArCH (Architectural Cultural Heritage) (Matrone et al., 2020b) and now part of the state-of-the-art since it has been published and made available for the scientific community (www.archdataset.polito.it). It consists of 15 point clouds for the DNNs training phases and 2 for the tests, for a total of 136 million labelled points. These scenes represent architectural assets and, in some cases, are also part of UNESCO sites. In particular, they represent different epochs and architectural styles. Figure 1 shows how it encompasses architectures from the early Middle Ages to the Romanesque, up to the Mannerism period and the Renaissance and Baroque ones. It also ranges from the European continent (Italy and France) to the Asian one (Indonesia). These scenarios certainly do not define an all-inclusive dataset, however, they constitute a good starting point for a solid state-of-art and ensure a good level of generalisability of the obtained results. The point clouds have been labelled manually, in order to provide a secure reference, not being deceptive and misleading for the DNNs.

2.2. ML and DL approaches in the CH domain

Within the CH domain, ML and DL techniques have been applied not only to the architectural field but also to art and archaeology.

In the field of archaeology and remote sensing, Recurrent-CNNs have been used for the identification of sites under the ground surface relying on LiDAR or shapefile data (Verschoof-van der Vaart and Lambers, 2019; Sharafi et al., 2016). A Google application named Fabricius, has been launched for the automatic translation of hieroglyphs (Chadwick, 2020) and so on. Other works attempting at classifying DCH images with different techniques are (Mathias et al., 2011; Oses et al., 2014; Llamas et al., 2017; Stathopoulou and Remondino, 2019), but they have still not exploited for 3D data as point clouds.

Starting from these studies, *semi-supervised* approaches have also been developed. Exploiting the deep NNs, they are particularly efficient for the CH domain, as they need a small portion of annotated data, overcoming the problem of lack of datasets. An example of this approach is the work of (Baraldi et al., 2018) in which learning of visual semantic embeddings have been investigated to provide an automatic annotation of historical document illustrations and captions. Both supervised and semi-supervised approaches have been tested. The comparison of visual and textual data is conducted through the creation of a shared embedding space, where the features can be compared based on distance. This semi-automatic approach is based on Maximum Mean Discrepancy (MMD) (Yan et al., 2017) in which the reproducing kernel Hilbert space is exploited to compare the distance between the expected results of the two distributions. In particular, this specific type of Hilbert space allows determining whether two functions are pointwise close (if they are close in the norm, they are also pointwise close) and its combination with the proposed weighted MMD is extremely useful for domain adaptation. VGG-19, a renowned CNN developed by the Visual Geometry Group of the University of Oxford with 19 layers (Simonyan and Zisserman, 2014), and ResNet-152, a residual deep network with 152 layers (He et al., 2016), have been chosen to encode input images. In this case, it is demonstrated how cross-domain reference is possible and that it is unnecessary to have a large amount of data as input. However, the use of 2D images partially simplifies the use of a DL framework, since Convolutional NNs (CNN) can be applied. This is more challenging with point clouds, because they are unordered and unstructured geometric data, so CNNs cannot be easily applied to them. In this case, three approaches have been developed for the semantic segmentation (Xie et al., 2019): i) *multiview-based* in which a set of images is created from the point cloud, ii) *voxel-based* where the cloud is rasterised in order to make it possible the application of CNNs, and finally, iii) the *point-based* methods in which the raw point cloud is directly consumed and semantic segmentation tasks are carried out by applying features-based approaches.

To the best of our knowledge, there are still few studies related to the topic addressed in this contribution. Some researchers exploit point clouds for semi-automatic or automatic elements recognition (Murtiyoso & Grussenmeyer, 2019a, 2019b) and the consequent reconstruction of BIM models (Bassier et al., 2020). Nevertheless, although with excellent results, they do not yet involve the use of DNNs. A closer work is the one of (Terruggi et al., 2020), which performs a semantic segmentation of heritage point clouds, with a good level of detail (architectural and decorative elements). Although, the use of DNNs, in this case, is not contemplated. The research is based on the use of 3D features (Weinmann et al., 2015), namely shape descriptors derived from a compound of eigenvalues ($\lambda_1 > \lambda_2 > \lambda_3$) obtained from the covariance matrix, able to describe and emphasise in a particularly explicit way the different architectural elements (Grilli & Remondino, 2020). These 3D features are used as a starting point for the RF classifier and the results are very promising. In (Grilli et al., 2019a) this approach is compared with the performances of some state-of-the-art DNNs, however, the chosen networks (1D/2D CNNs and a Bi-Long short-term memory Recurrent NN) are not suitable for the

exploitation of point clouds, thus leading to poor results. Starting from the just mentioned research, a comparison between the ML and DL approaches was developed in (Matrone et al., 2020a), highlighting the potentialities and criticalities of both methods, and demonstrating how they can be a viable path for the DCH. In particular, the network used in this last work is the DGCNN (Wang et al., 2019), based on a specific module called EdgeConv. This module captures local geometric structure while maintaining permutation invariance. It generates edge features that describe the relationship between a point and its neighbours instead of generating points' feature directly from embedding. DGCNN elaborates, in fact, a dynamic graph, i.e. it recomputes the graph in the feature space produced by each layer using nearest neighbours. So at each layer, the graph is updated with the nearest neighbours using the current feature space. In this paper, this network's adaptation, specially designed for the CH domain and called DGCNN-Modified (DGCNN-Mod) (Pierdicca et al., 2020), will be tested, evaluated, and improved. It allows exploiting, in addition to spatial coordinates, different 3D features coming from point clouds, in order to guide the *k*-NN approach in the selection of points neighbourhoods. This approach permits us to learn more discriminating features for the various classes of scenes.

3. Methodology

Starting from the results obtained in (Pierdicca et al., 2020), the methodology proposed (Figure 2) examines whether data augmentation techniques applied to a particular dataset, as the ArCH one, can be useful and effective as in the case of 2D datasets. Besides, an investigation on how to make the whole workflow more functional and "friendly" for external users has been carried out too. Firstly, a form of data augmentation is hence presented. Subsequently, a fine-tuning approach is proposed to understand if, also in the CH domain, it can lead to performances improvement, introducing a new scene in a pre-trained network. In fact, the peculiarities of each scene do not guarantee certain and definite results, as for other domains. This section is divided into two subsections: classic fine-tuning and fine-tuning with the addition of the RF classifier in the final part of the prediction have been both tested. In the latter case, the choice of adding the RF is due to the results obtained in (Grilli et al., 2019b), which have shown that in a short time and even in the presence of relatively limited data, it is able to provide excellent results.

3.1. Deep learning with the modified DGCNN (DGCNN-Mod+3Dfeat)

As said before, the DL approach adopted in this paper is based on the DGCNN-Mod network. This approach has been designed to consider in the first *k*-NN phase, not only the coordinates of normalised points but also other features like colour features transformations (Hue Saturation Value or Red Green Blue channels) and normal vectors. In this way, the *k*-NN method is aided in better learning neighbourhood points that allow generating more discriminating features. This approach has been further improved in (Matrone et al., 2020a), where additional input features have been added, leading to better results for semantic segmentation of point clouds in the CH domain. The new input features are 3D features based on ML approaches, so they are handcrafted. The modified network has been renamed DGCNN-Mod+3Dfeat, but it will be referred to "modified DGCNN" for the sake of simplicity in this article.

3.1.1. Data augmentation

Generally, a DL approach can be improved by using particular data augmentation techniques on the training data. In our case, we needed methods that can be applied to point clouds.

So, we have implemented five different techniques (Figure 3**)**:

- *rotation*, with random steps of 90 degrees;
- *clipping*, on random portions of the data;
- *spatial shifting*, on X and Y directions;
- *jittering*, by adding Gaussian noise on the data;
- *scaling*, by using a random scale factor, between a minimum and a maximum value.

These techniques are applied on the blocks (1x1 m with endless height) of the scenes that are fed into the network. At each epoch, for each block, one of these

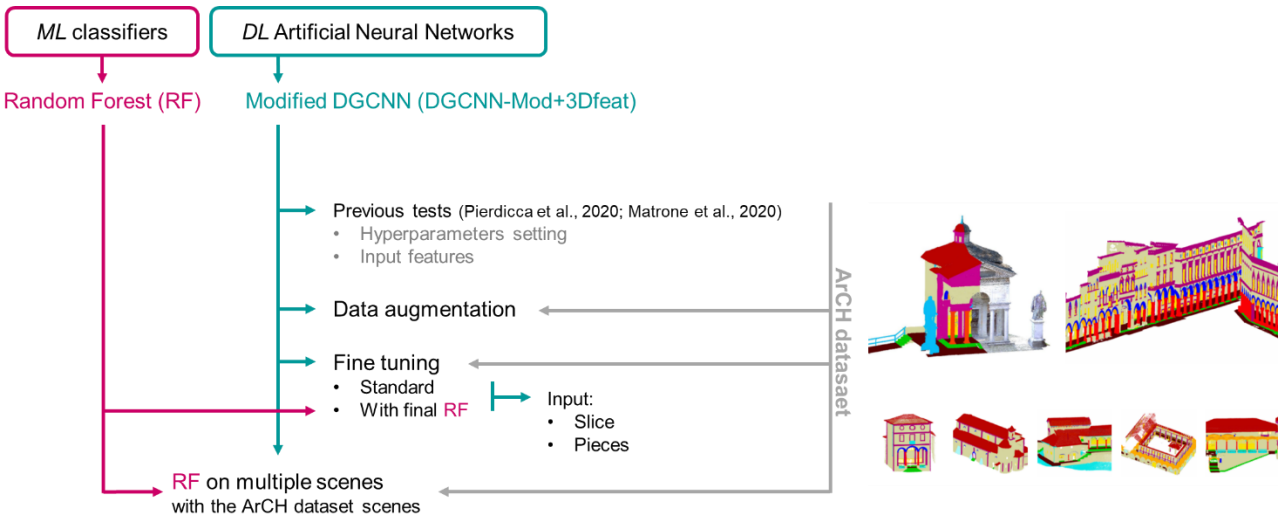
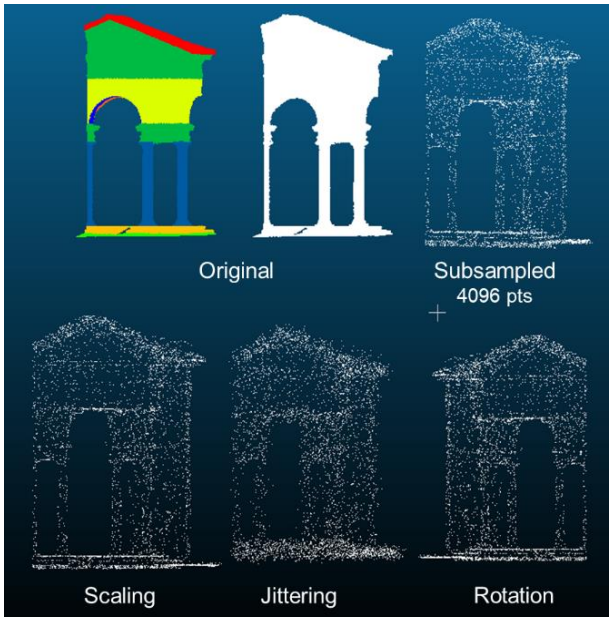
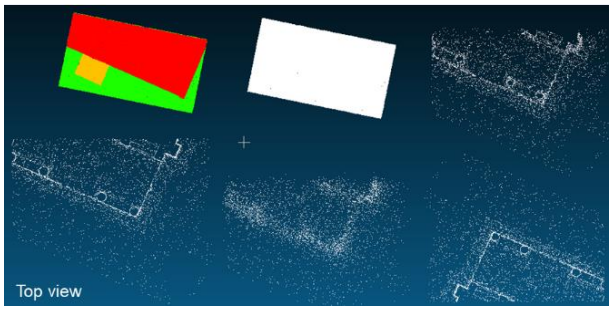


Figure 2: Research workflow and tests performed

methods is applied randomly. The approach is very similar to the one used in PointNet training (Qi et al., 2017), where the point cloud is augmented on-the-fly.



(a)



(b)

Figure 3: Example of data augmentation techniques. For representation purposes, a block of 10 x 10 m has been set: a) front view; b) top view.

3.2. Fine-tuning

In addition to data augmentation techniques, there is another method to improve the neural network's performance. This technique is called fine-tuning and allows the use of the weights from a pre-trained network, and re-train it on a new dataset. In this way, we can obtain good performance in a short time, compared to training from scratch. It is, therefore, a transfer learning approach. In our case, we want to use a network trained on the ArCH dataset scenes, then re-train it on a portion of a scene never seen by the network, and finally test it on the remaining portion. The basic idea is that, by quickly re-training the network on a few data of a new scene, the modified DGCNN could be oriented to better discriminate this new scene's classes.

3.2.1. Fine-tuning with NN features into RF

A hybrid approach has also been tested: a fine-tuning with the addition of the RF classifier in the final part of the prediction phase. The choice of adding the RF is due to the results obtained in (Grilli et al., 2019b), which have shown that in a short time and even in the presence of relatively limited data, it is able to provide excellent results.

As in the classic fine-tuning technique, the network weights, pre-trained on the ArCH dataset scenes, have been employed. Then, the final part of the modified DGCNN performing the segmentation of the points is excluded. In this way, the network will be used as a *feature extractor* method.

In the second phase, a scene of the dataset never seen by the network is chosen: this scene is divided into one part for training and one for the test. Afterwards, the features of both parts are extracted using the *feature extractor*, and exploited as input for training the RF classifier.

3.3. RF trained on multiple scenes

A training of only the RF classifier was also carried out using the original features and scenes of the ArCH dataset to obtain a complete and adequate comparison of the various methods. Besides, in Section 4, a fine-tuning of the classifier hyperparameters is performed, such as number and depth of trees or the choice of *gini* and *entropy* measures for node impurities. These are essential elements to achieving better performances for the RF.

The classifier was trained on the same scenes involved in the other methods in order to obtain congruent results. An identical approach was maintained for the testing phase too, where the same portion of the scene is used in all the compared approaches.

4. Results and discussions

This section reports the results of the tests conducted according to the methodology described above. The tests proposed in Section 4.1 concern data augmentation. Only the best results are summarised. Section 4.2, on the other hand, focuses on fine-tuning, divided into standard configuration and with the addition of RF. Finally, the results of tests conducted with only the RF trained on multiple scenes of the ArCH dataset are reported in Section 4.3.

The performances are shown in terms of 3 different metrics: the Overall Accuracy (OA) of the predicted points, the F1-Score for the individual classes and its Weighted Average (WAvg) value. All experiments have been implemented using the Tensorflow framework and the Python 3 language. The network fine-tuning technique was performed by lowering the learning rate of the original training by 1/10, and the SGD (Stochastic Gradient Descent) technique has been set as an optimiser. As stated in (Matrone et al., 2020a), the *scaler2* pre-processing technique, implemented through the Scikit-Learn library, is used for data normalisation since it proved to be the best method. Compared to *scaler1*, which standardises features by removing the mean and scaling to unit variance, *scaler2* removes the median and scales the data according to the quartile range, becoming more robust to outliers. In addition, other specific techniques have been tested, such as the focal loss function and skip connections. The *focal loss* is a particular function designed to solve issues due to unbalanced datasets. We introduced it because in the ArCH dataset some classes have fewer points than others (e.g. Wall and Roof compared to Columns or

Door/Window). Instead, the *skip connection* is a particular technique that allows concatenating the input features of a network with those learned in the last layers, to improve the model convergence.

4.1. Data augmentation

For this group of experiments, standard data augmentation parameters have been set up, but other tests are ongoing with different possible configurations. In particular, the following parameters have been chosen: rotation of 90 degrees, 0.06 for jittering standard deviation, 0.18 for the clipping factor, 0.1 spatial shifting factor, scale factor between 0.8 and 1.25 values.

The results (Table 1) show that no marked improvement is achieved if comparing the only overall accuracy (OA), but specific considerations can be made on the single classes.

Table 1. Comparison of results: F1-score for each class, WAvg and OA of the data augmentation tests

	Reference tests		*Data augmentation*		
Focal loss	Yes	Yes	No	Yes	Yes
Skip connect	No	Yes	No	No	Yes
Arch	0.08	0.05	0.02	0.00	0.05
Column	0.53	0.40	0.31	0.16	0.26
Moulding	0.37	0.43	0.37	0.41	0.46
Floor	0.83	0.81	0.81	0.79	0.81
Door/Window	0.39	0.41	0.56	0.47	0.50
Wall	0.84	0.84	0.85	0.84	0.84
Stair	0.83	0.80	0.79	0.77	0.81
Vault	0.85	0.85	0.88	0.85	0.87
Roof	0.95	0.95	0.96	0.96	0.96
WAvg	0.83	0.84	0.84	0.83	0.84
OA	0.84	0.85	0.85	0.85	0.85

The classes with more points in the training set, such as Floor, Wall or Roof, remain almost unchanged. A slight improvement is noted for the Vaults and Mouldings in a single configuration. On the other hand, the Door/Window class improvement is significant, where an average of 0.51 is registered in the data augmentation, compared to the 0.4 average of the reference tests. The opposite behaviour is registered for the Arch and Column classes where, especially in the latter, the results are better without the data augmentation. In particular, for the Columns, the decline in performance is significant and confirms the results of a further test carried out in which scenes were added with only columns apart (Figure 4).

The last test results may be due to the introduction of this kind of scenes, which led the network to learn that columns must be far away from any other object, except for the floor. So, if the network is then tested with scenes having columns near any other objects (arches, vaults, mouldings and so on), it will probably not recognise them.

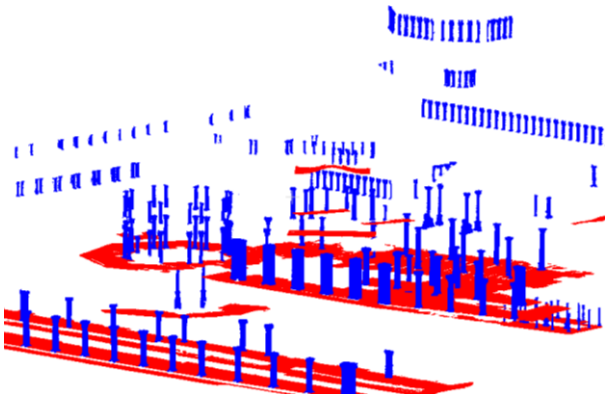


Figure 4*:* Example of a scene introduced in the training set for data augmentation of the column class.

4.2. Fine-tuning

The tests with the standard fine-tuning have been further divided into two sub-groups. In the first, starting from the pre-trained network, some “pieces” of the test scene are given as input, including all the possible classes, while in the second, a “slice” of the scene is cut (Figure 5). In both configurations, the test is performed on the complementary scene (RGB data in Figure 5).

The first option was defined on the basis of tests conducted in (Matrone et al., 2020a) in which parts of each class are annotated and given as input to the RF. The second, on the other hand, was dictated by the results obtained in the data augmentation tests. In that case, the addition of scenes with only columns as input for training, led to a net worsening of performances, probably due to the inability of the network (based on graphs) to determine relationships with neighbouring points.

In all the tests (Table 2) the network has been pre-trained with the hyperparameters that guaranteed the best result (scaler 2 with only focal loss) in the previous tests, while the fine-tuning was conducted by varying the scalers (1, 2 or none) and adding or removing both focal loss and skip connections.

The results show how, within the fine-tuning, the OA is better if a “slice” of the scene is provided as input. This outcome confirms what already emerged from the data augmentation: connections matter. The network learns better the connections between the various classes of objects only if congruent and complete scenes are given for its training phase. It demonstrates that the modified DGCNN is learning the spatial relations among the various architectural elements.

Scenes containing holes between object connections tend to mislead the network into learning the wrong discriminating features.

If compared with the reference tests, the results validate the effectiveness of fine-tuning, especially for those classes in which there are fewer points in the training set and the elements to be recognised are more heterogeneous. Indeed, the columns, arches, doors/ windows, and mouldings are very different from each other within the same class. This is a peculiarity of built heritage, which is in contrast with the basic functioning of the neural network: the more it sees an element of the same type, the more it will be able to recognise it during the final prediction.

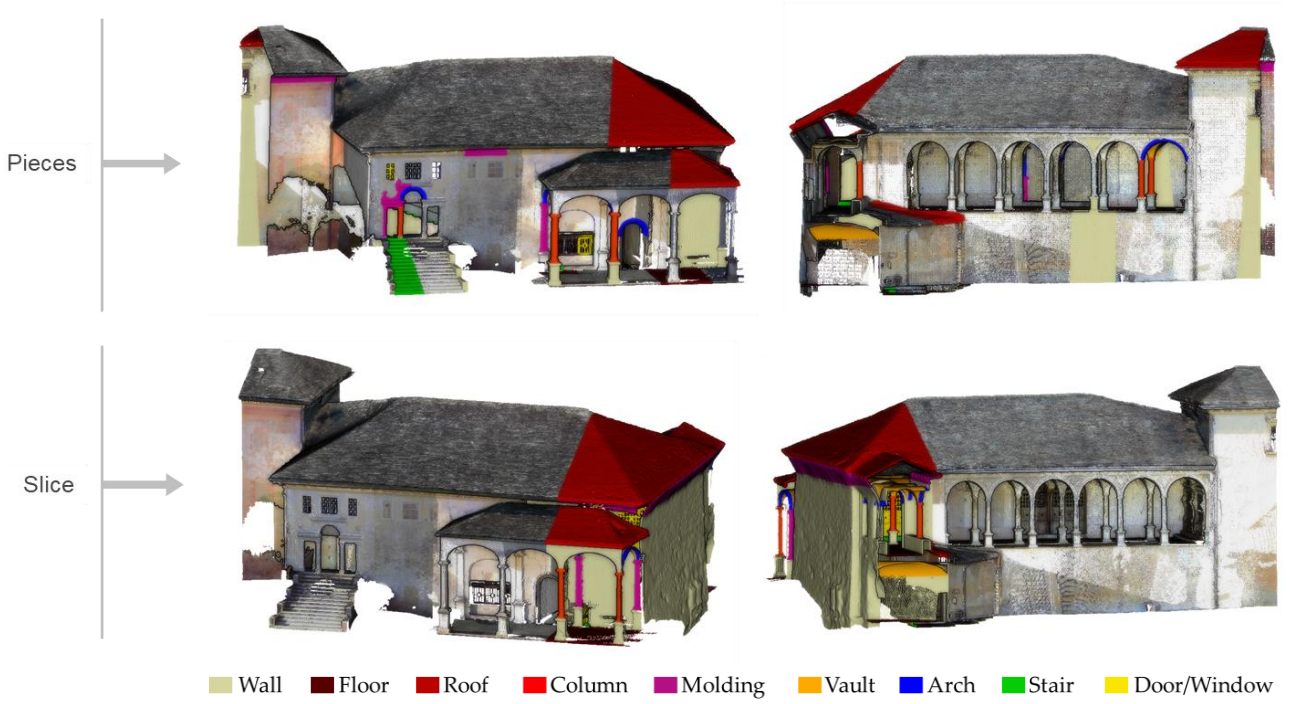


Figure 5. Subdivision of the test scene for the fine-tuning experiments. In the first case, only the coloured pieces have been used to fine-tune the network, while the RGB complementary part constitutes the test set. In the second case, an entire slice of the scene is used as input for the fine-tuning. The different colours correspond to the various classes.

Table 2. Comparison of results: F1-score for each class, WAvg and OA of fine-tuning tests.

	Reference tests		*Fine-tuning*		
			Slice		Pieces
Focal loss	Yes	Yes	Yes	Yes	Yes
Skip connect.	No	Yes	No	Yes	No
Arch	0.08	0.05	0.24	0.35	0.15
Column	0.53	0.40	0.81	0.65	0.63
Moulding	0.37	0.43	0.54	0.55	0.31
Floor	0.83	0.81	0.63	0.49	0.81
Door/Window	0.39	0.41	0.66	0.60	0.01
Wall	0.84	0.84	0.83	0.80	0.54
Stair	0.83	0.80	0.06	0.06	0.77
Vault	0.85	0.85	0.87	0.90	0.78
Roof	0.95	0.95	0.97	0.97	0.91
WAvg	0.83	0.84	0.81	0.80	0.70
OA	0.84	0.85	0.84	0.81	0.71

By fine-tuning part of the scene that will be used as a test, the network trains itself specifically on the same type of architectural elements that it will then find in the complementary test set, overcoming the aforementioned issue. For the classes that, instead, have a greater number of points and have more standard elements, such as walls or roofs, the value remains almost unchanged. A separate discussion should be made for the Stair class where the total absence of the element in the input scene could have negatively affected the results.

4.2.1. Fine-tuning with NN features into RF

Since RF proved to be an excellent classifier for the semantic segmentation task of built heritage (Grilli et al., 2019b), an attempt to extract the features learned from the network and give them as input to the RF has been carried out. This procedure would allow, starting from a pre-trained DNN, to save the features in a separate set and use them whenever necessary, to directly train the RF. In this way, the user would not even have to annotate a small part of the test scene (Grilli et al., 2019b; Matrone et al., 2020a).

Likewise Section 4.2, the experiments were divided according to the input used for the RF: “slice” or “pieces”.

Several configurations were tested:

- Scaler 1, 2 or none for the DGCNN training;
- Both gini and entropy as measures of the impurity of a node;
- 100, 150 or 200 for the number of trees. It was noticed that over 150, the performances began to decay, therefore no tests were carried out beyond 200;
- 10, 20 or 50 and none for the depth. After noting that the choice of none guaranteed better results, causing just a slightly higher computational time, none was chosen for all subsequent tests.

Since these tests were performed by using the output of the NN as input for the RF, in Table 3 the references are the best result obtained with the DGCNN-Mod, and those obtained with only the RF trained on pieces of the test scene.

Table 3. Comparison of results: F1-score for each class, WAvg and OA of fine-tuning + RF tests.

	Reference tests		*Fine-tuning*		
			Slice		Pieces
	DNN	RF	Features from NN into RF		
Scaler	2	-	-	-	2
N. of trees		100	200	150	150
Meas. impurity		Gini	Gini	Entropy	Gini
Arch	0.05	0.46	0.28	0.27	0.10
Column	0.40	0.91	0.24	0.31	0.68
Moulding	0.43	0.55	0.61	0.63	0.17
Floor	0.81	0.94	0.56	0.56	0.75
Door/Window	0.41	0.32	0.32	0.34	0.01
Wall	0.84	0.87	0.81	0.81	0.52
Stair	0.80	0.82	0.07	0.08	0.77
Vault	0.85	0.90	0.90	0.90	0.73
Roof	0.95	0.87	0.95	0.96	0.89
WAvg	0.84	0.85	0.79	0.79	0.68
OA	0.85	0.84	0.81	0.82	0.70

The only results that show an improvement w.r.t. both the DNN and the RF have been highlighted.

In general, this approach does not seem to bring actual benefits, but in terms of OA, it nevertheless confirms the achievement of similar performances, in the case of “slice”, to those of the reference tests. This outcome demonstrates how, on the one hand, the features learned from the network are really able to describe the classes on which it has been trained and, on the other hand, how the addition of the RF classifier does not necessarily guarantee better results. It is the modality with which it is trained that mainly affects, not the mere prediction task.

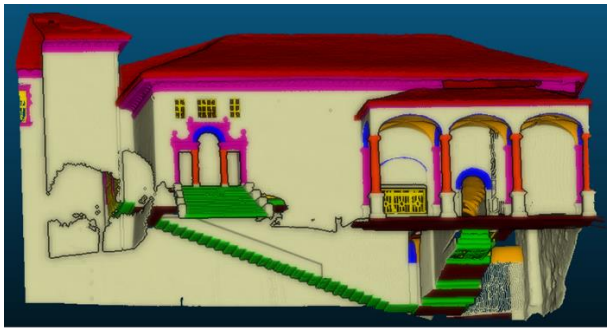
4.3. RF trained on multiple scenes

On the basis of the previous outcomes, an attempt was also made by training the RF with the same scenes from the ArCH dataset used for training the DGCNN-Mod network. The parameters chosen for the training phase are the same of those selected for the tests in Section 4.2.1. However, the results obtained with different types of configurations have not achieved a sufficient level of performance to constitute a valid reference for the state-of-the-art (Table 4). Their OA ranges from 0.15 to 0.68, showing a strong dependence on the type of scaler used for the training set: scaler 2 ranges from 0.15 to 0.18, scaler 1 is in the range of 0.32 to 0.40 and the use of no scaler led to 0.65-0.68.

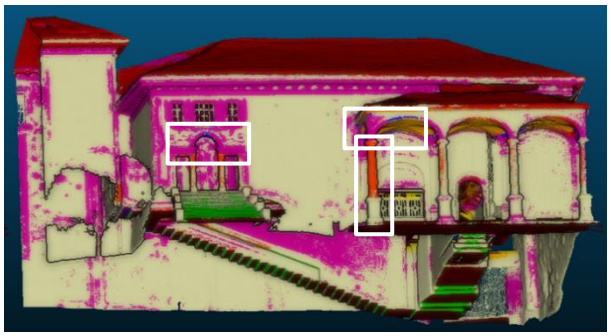
Table 4: Best results of the RF trained on multiple scenes.

	Measure of impurity	*Scaler*	*Best OA*
RF 1	Gini	-	0.663
RF 2	Entropy	-	0.678
RF 3	Gini	1	0.367
RF 4	Entropy	1	0.397
RF 5	Gini	2	0.176
RF 6	Entropy	2	0.184

From the visual comparison (Figure 6) and the analysis of the metrics of the individual classes, it is clear that the predominant class that has been misclassified is Moulding. However, even if the result does not achieve the performances of the other tested methods, some positive elements can still be noted, including (in the white rectangles) the recognition of some parts of the arches, not recognised in the DGCNN-based methods, as well as an entire column correctly labelled, with the exclusion of the base. Nonetheless, the result was not considered sufficient to deepen this approach.



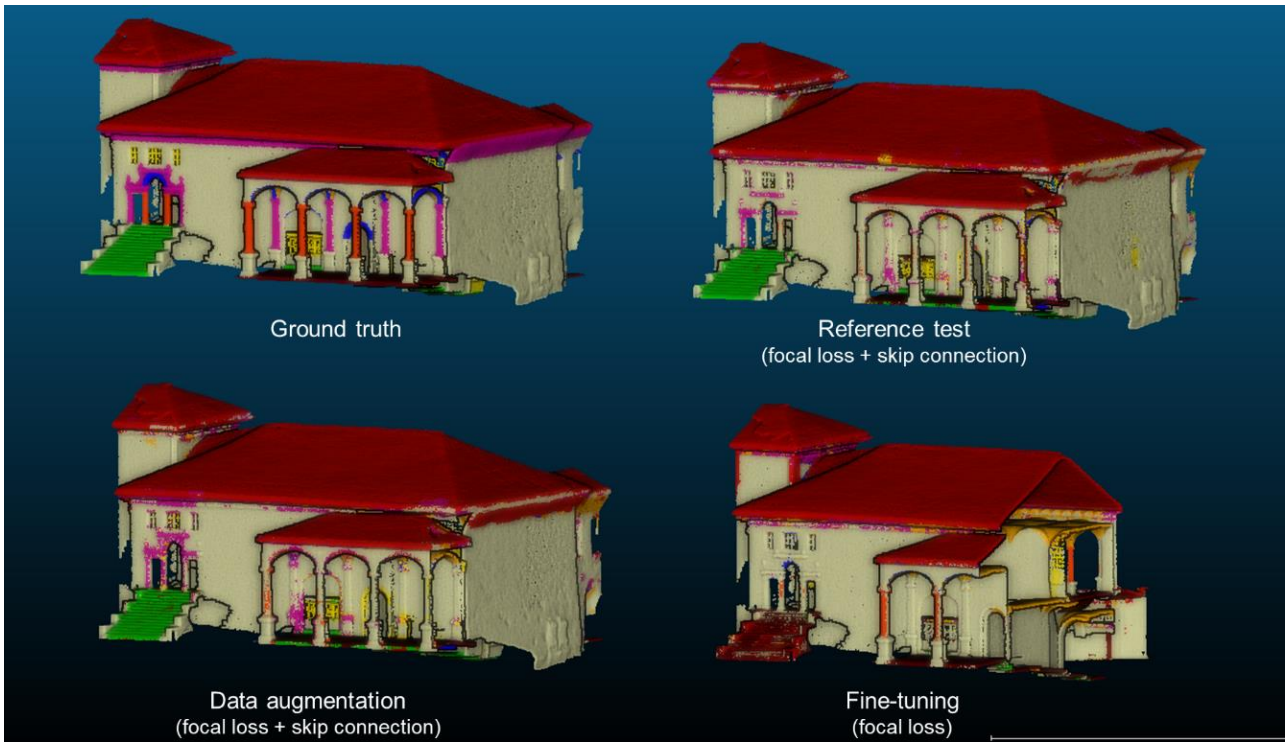
(a)



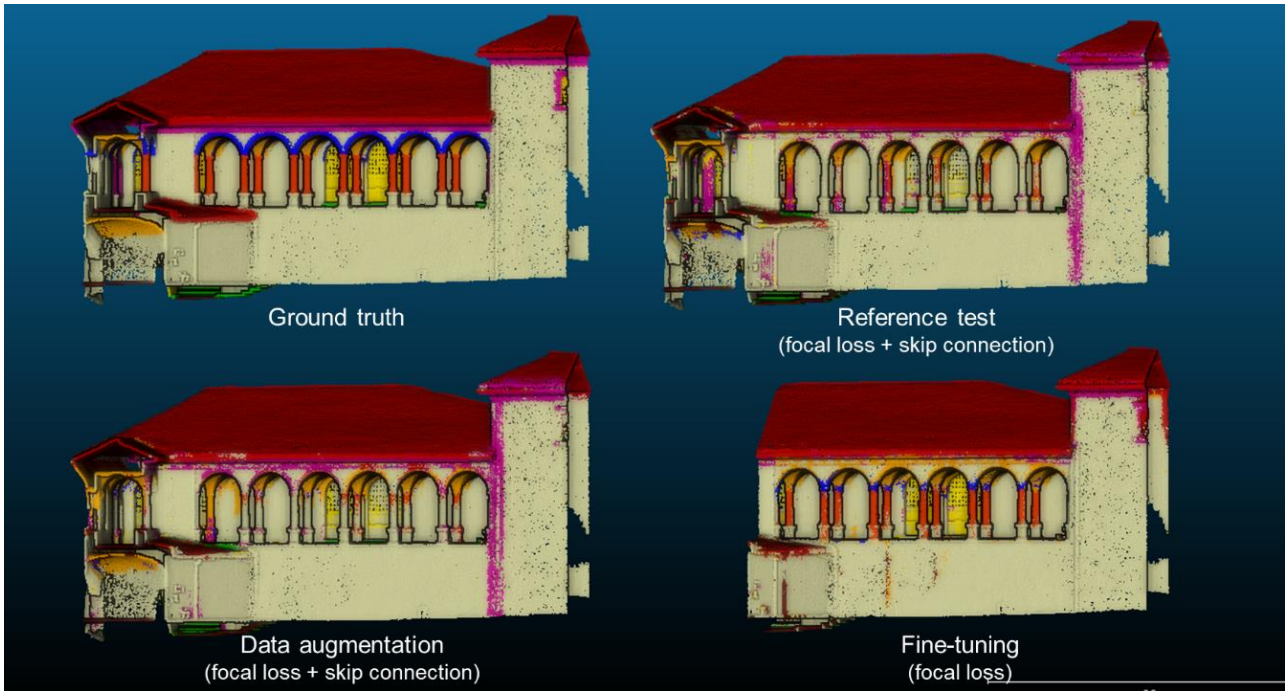
(b)

Figure 6: Visual inspection of the results with the RF trained on multiple scenes: a) ground truth; b) predicted scene.

These results show that RF is not suitable to segment objects from totally different scenes (diverse styles and geometries) using directly the original features. It needs to have: i) more discriminating features, coming for example from other methods such as pretrained DNNs (e.g. the modified DGCNN herein presented, utilised as a *feature extractor*), or ii) geometric features similar to those found in the test scene. This last assumption is confirmed by the works of (Grilli et al., 2019b; Teruggi et al., 2020) where the classifier has achieved excellent and very promising results being both trained and tested with parts of the same scene. In this way, the input features are very similar to those the algorithm will encounter when classifying and predicting the remaining part of the scene, improving the final results.



(a)



(b)

Wall Floor Roof Column Molding Vault Arch Stair Door/Window

Figure 7: Ground truth and predictions of the test scene (best result for each method): a) south façade; b) north façade.

5. Conclusions

In this paper, a new approach for the semantic segmentation of heritage point clouds is presented. Starting from previous tests, close attention is paid to alternative methods to improve performances and try to take advantage of pre-trained networks to speed up and simplify their use for external users.

Tests conducted on data augmentation have shown that they do not affect overall performances, but still provide proper support for those classes with fewer points, especially if associated with focal loss.

The tests on the NN fine-tuning have instead given rise to multiple considerations. Firstly, the standard fine-tuning is able to achieve performances almost equal to those where only the modified DGCNN is used. Therefore, they confirm that, once the DNN is pre-trained, data processing and prediction times can be significantly reduced (from about 48 h to just over 0.5 h), in the case of heritage point clouds too. As regards the

use of the modified DGCNN as a *feature extractor* and the RF as a classifier, the achievement of performances similar to the reference tests is obtained, with some classes even better detected (Mouldings). However, there is a strong dependence of the classifier on the type of scaler used to normalise the data. The absence of the scaler guarantees better results. On the other hand, the measures of impurity of nodes do not significantly affect the results. The training of the RF on several scenes does not lead to good performances, thus not proving to be a real alternative to the methodology here proposed. Except for some cases, it has not been possible to identify a common and unique pattern able to define precise guidelines for the hyperparameters to be set in the fine-tuning or data augmentation tests (Figure 7).

In conclusion, to answer the initial research questions, it is possible to use DL techniques also in the CH domain and, specifically, of built heritage point clouds. The modified DGCNN has proven to achieve performances similar to those of the more consolidated ML classifiers. Besides, it guarantees the possibility to avoid manual annotation by the end-user, if fine-tuning is not carried out, but the weights saved by the pre-trained network are directly used to make the prediction. With regards to the use and exploitation of this methodology by external users, it can be stated that: from the point of view of the required computational times and resources, they can be significantly reduced thanks to the possibility of pre-training the network and then use both the extracted features and the weights for subsequent tests. Users can, in fact, label only a small part of the new test scene and then rely on the data coming from the pre-trained network. Moreover, since the categories have been already defined and the point clouds have been manually labelled by a domain expert, even non-expert users can profit from the methodology. Finally, the original code has been implemented and adequately generalised for other datasets, so it is unnecessary to deeply intervene on the algorithms and only a few precautions are required to make it fully operating and exploitable. These latter elements make it relatively user-friendly if compared to other DL approaches. In this way, even non-programming users or domain experts are facilitated and can take advantage of this methodology.

Future developments of this research are going to be oriented towards the interpretability and explicability of the modified DGCNN and the use of taxonomies or ontologies to guide the learning of the network.

References

Armeni, I., Sener, O., Zamir, A. R., Jiang, H., Brilakis, I., Fischer, M., & Savarese, S. (2016). 3D semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1534–1543. https://doi.org/10.1109/CVPR.2016.170

Baraldi, L., Cornia, M., Grana, C., & Cucchiara, R. (2018). Aligning text and document illustrations: towards visually explainable digital humanities. In 24th International Conference on Pattern Recognition (ICPR), 1097–1102. IEEE. https://doi.org/10.1109/ICPR.2018.8545064

Bassier, M., Yousefzadeh, M., & Vergauwen, M. (2020). Comparison of 2D and 3D wall reconstruction algorithms from point cloud data for as-built BIM. *Journal of Information Technology in Construction (ITcon)*, *25*(11), 173–192. https://doi.org/10.36680/j.itcon.2020.011

Boulch, A., Guerry, J., Le Saux, B., & Audebert, N. (2018). SnapNet: 3D point cloud semantic labeling with 2D deep segmentation networks. *Computers & Graphics*, *71*, 189–198. https://doi.org/10.1016/j.cag.2017.11.010

Chadwick, J., (2020). Google launches hieroglyphics translator that uses AI to decipher images of Ancient Egyptian script. Available at https://www.dailymail.co.uk/sciencetech/article-8540329/Google-launches-hieroglyphics-translator-uses-AI-decipher-Ancient-Egyptian-script.html Last access 24/11/2020

Fiorucci, M., Khoroshiltseva, M., Pontil, M., Traviglia, A., Del Bue, A., & James, S. (2020). Machine learning for cultural heritage: a survey. *Pattern Recognition Letters*, *133*, 102–108. https://doi.org/10.1016/j.patrec.2020.02.017

Geiger, A., Lenz, P., Stiller, C., & Urtasun, R. (2013). Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research*, *32*(11), 1231–1237. https://doi.org/10.1177/0278364913491297

Grilli, E., & Remondino, F. (2019). Classification of 3D digital heritage. *Remote Sensing*, *11*(7), 847. https://doi.org/10.3390/rs11070847

Grilli, E., & Remondino, F. (2020). Machine learning generalisation across different 3D architectural heritage. *ISPRS International Journal of Geo-Information*, *9*(6), 379. https://doi.org/10.3390/ijgi9060379

Grilli, E., Özdemir, E., & Remondino, F. (2019a). Application of machine and deep learning strategies for the classification of heritage point clouds. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences,* XLII-4/W18, 447–454. https://doi.org/10.5194/isprs-archives-XLII-4-W18-447-2019

Grilli, E., Farella, E. M., Torresani, A., & Remondino, F. (2019b). Geometric features analysis for the classification of cultural heritage point clouds. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences,* XLII-2/W15, 541–548. https://doi.org/10.5194/isprs-archives-XLII-2-W15-541-2019

Hackel, T., Savinov, N., Ladicky, L., Wegner, J. D., Schindler, K., & Pollefeys, M. (2017). Semantic3d.net: A new large-scale point cloud classification benchmark. *arXiv:1704.03847*

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778. *arXiv:1512.03385*

Korc, F., & Förstner, W. (2009). eTRIMS Image Database for interpreting images of man-made scenes. *Dept. of Photogrammetry, University of Bonn, Tech. Rep. TR-IGG-P-2009-01*.

Landrieu, L., & Simonovsky, M. (2018). Large-scale point cloud semantic segmentation with superpoint graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4558–4567. *arXiv:1711.09869*

Llamas, J., M Lerones, P., Medina, R., Zalama, E., & Gómez-García-Bermejo, J. (2017). Classification of architectural heritage images using deep learning techniques. *Applied Sciences*, *7*(10), 992. https://doi.org/10.3390/app7100992

Mathias, M., Martinovic, A., Weissenberg, J., Haegler, S., & Van Gool, L. (2011). Automatic architectural style recognition. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences,* XXXVIII-5/W16, 171–176. https://doi.org/10.3390/app7100992

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

Matrone, F., Lingua, A., Pierdicca, R., Malinverni, E. S., Paolanti, M., Grilli, E., Remondino, F., Murtiyoso, A., & Landes, T. (2020b). A benchmark for large-scale heritage point cloud semantic segmentation. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B2-2020, 1419–1426. https://doi.org/10.5194/isprs-archives-XLIII-B2-2020-1419-2020

Murtiyoso, A., & Grussenmeyer, P. (2019a). Automatic heritage building point cloud segmentation and classification using geometrical rules. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences,* XLII-2/W15, 821–827. https://doi.org/10.5194/isprs-archives-XLII-2-W15-821-2019

Murtiyoso, A., & Grussenmeyer, P. (2019b). Point cloud segmentation and semantic annotation aided by GIS data for heritage complexes. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W9, 523–528. https://doi.org/10.5194/isprs-archives-XLII-2-W9-523-2019

Oses, N., Dornaika, F., & Moujahid, A. (2014). Image-based delineation and classification of built heritage masonry. Remote Sensing, *6*(3), 1863–1889. https://doi.org/10.3390/rs6031863

Park, Y., & Guldmann, J. M. (2019). Creating 3D city models with building footprints and LIDAR point cloud classification: A machine learning approach. *Computers, Environment and Urban Systems*, *75*, 76–89. https://doi.org/10.1016/j.compenvurbsys.2019.01.004

Pierdicca, R., Paolanti, M., Matrone, F., Martini, M., Morbidoni, C., Malinverni, E. S. & Lingua, A. M. (2020). Point cloud semantic segmentation using a deep learning framework for cultural heritage. *Remote Sensing*, *12*(6), 1005. https://doi.org/10.3390/rs12061005

Qi, C. R., Su, H., Mo, K., & Guibas, L. J. (2017). Pointnet: Deep learning on point sets for 3D classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 652–660. *arXiv:1612.00593*

Sharafi, S., Fouladvand, S., Simpson, I., & Alvarez, J. A. B. (2016). Application of pattern recognition in detection of buried archaeological sites based on analysing environmental variables, Khorramabad Plain, West Iran. *Journal of Archaeological Science: Reports*, 8, 206–215. https://doi.org/10.1016/j.jasrep.2016.06.024

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*

Stathopoulou, E. K., & Remondino, F. (2019). Semantic photogrammetry: boosting image-based 3D reconstruction with semantic labeling. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *42*(2), W9. https://doi.org/10.5194/isprs-archives-XLII-2-W9-685-2019

Teboul, O., Kokkinos, I., Simon, L., Koutsourakis, P., & Paragios, N. (2012). Parsing facades with shape grammars and reinforcement learning. *IEEE transactions on pattern analysis and machine intelligence*, *35*(7), 1744–1756. https://doi.org/10.1109/TPAMI.2012.252.

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 Sensing*, *12*(16), 2598. https://doi.org/10.3390/rs12162598

Tyleček, R., & Šára, R. (2013). Spatial pattern templates for recognition of objects with regular structure. In *German Conference on Pattern Recognition*, Springer, Berlin, Heidelberg, 364-374. https://doi.org/10.1007/978-3-642-40602-7_39

Verschoof-van der Vaart, W. B., & Lambers, K. (2019). Learning to Look at LiDAR: the use of R-CNN in the automated detection of archaeological objects in LiDAR data from the Netherlands. *Journal of Computer Applications in Archaeology*, *2*(1). https://doi.org/10.5334/jcaa.32

Wang, Y., Sun, Y., Liu, Z., Sarma, S. E., Bronstein, M. M., & Solomon, J. M. (2019). Dynamic graph CNN for learning on point clouds. *ACM Transactions on Graphics*, *38*(5), 1–12. *arXiv:1801.07829*

Weinmann, M., Jutzi, B., Hinz, S., & Mallet, C. (2015). Semantic point cloud interpretation based on optimal neighborhoods, relevant features and efficient classifiers. *ISPRS Journal of Photogrammetry and Remote Sensing*, *105*, 286–304. https://doi.org/10.1016/j.isprsjprs.2015.01.016

Xie, Y., Tian, J., & Zhu, X. X. (2019). Linking points with labels in 3D: a review of point cloud semantic segmentation. *arXiv:1908.08854*

Yan, H., Ding, Y., Li, P., Wang, Q., Xu, Y., & Zuo, W. (2017). Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2272–2281). *arXiv:1705.00609*