International Journal of Architectural Computing
©The Author(s) 2021
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
www.sagepub.com/
SAGE

An Artificial Neural Network Framework for annotating and classifying Architectural Structure and Style of Built Heritage in 3D

George Artopoulos[1], Maria I. Maslioukova[2], Christina Zavou[2], Melinos Averkiou[2,3], Andreas C. Andreou[2], Marissia Deligiorgi[1]

Abstract

The article presents an online interface based on Deep Neural Networks (DNNs) for identifying and comparing the structure and architectural style of buildings using 3D data. The research contributes a new approach to learning how to identify a building's structural components along with their stylistic influences, aiming to contribute new applications of Machine Learning (ML) that go beyond their use for accelerating the segmentation process of reality capture data (e.g., 3D point clouds) for automation in Historic Building Information Modeling (HBIM). Contributing to recent work in the literature, the research presented explores the use of neural networks trained on point cloud datasets generated by means of photogrammetry and aims to develop an integrated 3D interface for researchers and scholars in digital humanities.

Keywords

Deep Learning, Convolutional Neural Networks, Built Heritage, 3D reality capture of architecture, semantic annotation.

[1]The Cyprus Institute
[2]University of Cyprus
[3]CYENS - CENTRE OF EXCELLENCE

Corresponding author:
George Artopoulos, The Cyprus Institute
Email: g.artopoulos@cyi.ac.cy

1 Introduction

One of the fundamental classification operations in architectural historic research is the periodisation (Jiménez-Badillo et al. 2010). In architectural enquiries the chronological identification and construction phase analysis of a building poses many challenges. Oftentimes this operation requires on site observation; this is typically done by experts who classify artefacts chronologically based on spatial and social context, technique of production, provenance, style and geometric or material features (Baratin et al. 2012).

The article presents an online interface based on Artificial Intelligence techniques, more specifically, on 3D Convolutional Neural Networks (3D CNNs) and Support Vector Machines (SVMs), for identifying and comparing the structure and architectural style of buildings using 3D data. The research contributes a new approach in learning how to identify a building's architectural components (e.g., arch, dome), and their stylistic influences (e.g., Gothic, Byzantine). There is a longstanding effort in the application of analytical methods (i.e., based on logic operations) in architectural education for the study of the geometric and topological configuration of plans and facades of buildings (Mitchell et al. 1991; Gero 2020). These methods rely on 2D representations and abstractions of architecture, e.g., architectural drawings and floor plans.

In the last decades, progress in computer vision methods allowed the application of Deep Learning (DL) in style analysis (He et al. 2015; Yoshimura et al. 2018). Furthermore, until recently, most of the research in architectural style analysis was conducted based on architectural drawings, or floor plan configurations like the methods above (Duarte and Rocha 2006), but advances in computer vision have allowed researchers to handle more complex and bigger datasets, drawing from developments in other fields (Teboul et al. 2011), e.g., 2D image-based retrieval of information by means of CNNs ((Llamas et al. 2016); (Shalunts 2015)).

1.1 Segmentation operations for the classification of architectural representations

Comparative study of similitude and the differences that co-form a type is considered as a prerequisite for any generalisation procedure for classification purposes, and in the past this understanding inspired research in analytical studies of typological definitions of different historical periods (March and Steadman 2021). Distinct from this, the paper presents how, instead of resorting to logical operations of geometric transformation to infer the parameters of an archetypical (ideal) model - against which all variations of a type, or architectural style would be compared for similitude, the authors are exploring the capacities of training neural networks (NN) to quantify whether a building conforms with an architectural style by resembling by a certain percentage of 'truth,' i.e., by being compared to an annotated representation of an existing building (and not an 'ideal' type) that is considered the 'ground truth' of the style under question.

In addition to classification operations, another area of research this paper contributes to, regards the use of DL for building part segmentation. The advent of deep learning techniques is driving the experimental application of neural networks in the architecture and construction industry for processing and automating the segmentation of reality captured data, including applications in conservation studies (Chiabrando et al. 2017). This process is accelerated by the extensive penetration of Building Information Modelling (BIM) in architectural practice. BIM is a method that imposes the utilisation of standardised and parametricised building components, such as windows, doors, walls, floors (BIM families) in design representation. This process of standardised representation of building components is imposing limitations, especially in the case of heritage buildings, which often comprise of non-standard elements, or variations of parts (e.g., decorative elements, window, door and corner frames) – a challenge broadly discussed in the research area of HBIM applications (Macher et al. 2017; Wang et al. 2015).

The article brings theoretical considerations about the use of canonical expressions of a style for the valorisation of its variations (e.g., identified stylistic influences in Cypriot built heritage), together with practical aspects of implementing DL methods in architecture research for education. In addition, the article discusses about the performance of these methods in accelerating the segmentation process of reality captured data (e.g., 3D point clouds), an operation that is increasingly topical in literature occupied with automation of scan-to-BIM tasks.

The presented work is part of ANNFASS [1] (An Artificial Neural Network Framework for understanding historical monuments Architectural Structure and Style), a project funded by Research & Innovation Foundation [2]. The competitive Call through which ANNFASS was funded encourages local research and technological innovation, focusing on fields and societal challenges such as the safeguarding and promotion of the cultural heritage of Cyprus. ANNFASS responds to this challenge by developing an online platform and framework for digital humanities, that utilises ML for the classification of architectural elements and stylistic influences of Cypriot monuments.

1.2 Hybridisation of architectural styles: the case of the Cypriot architectural heritage

‘Archaeologists investigating a low hill on the southern edge of Nicosia in a millennium or two’s time are likely to discover the remains of a monumental structure with a bizarre mixture of architectural styles and motifs: Byzantine domes and column capitals, Gothic mouldings and windows, Venetian lions, Ottoman lattices, Cypriot vernacular arches, and British coats of arms [. . .].’ (Given 2005)

In the context of the geographical scope of the ANNFASS project in Cyprus, this section discusses Cypriot architectural heritage to exemplify the scope of using CNNs in annotating variations of a style. The history of architecture of Cyprus includes churches and mosques, authorities’ buildings, as well as vernacular houses, that were built under the Lusignan and Venetian rules from the twelfth to the sixteenth century, during the Ottoman empire and the early modern times. The majority of this architecture is the result of multi-period adaptations to the local climatic conditions and additions of architectural solutions previously developed in central Europe resulting in hybrid styles, including the so-called Neo-Gothic, Neo-Tudor/Elizabethan, New-Palladian, British ‘cottage vernacular’ (ibid. 38) (Given and Smith 2003; Schaar et al. 1995; Hamilakis 1996), some of which are studied here.

1.3 Representing architecture through geometric principles and the application of analytical rules in studying variation of architectural models

The early work on formal representations by Lionel March and Philip Steadman (March and Steadman 2021) was based on sets of geometric transformations (and mathematical functions) of groups of shapes. Informed by Group Theory, these studies applied a formal language through symmetry/symmetry-breaking rule-based hierarchical ordering of transformational functions, which led the way in researching shape-grammars (Stiny 1985). Logical research concerns various analysis systems (e.g. grammars) that first became known as shape grammarss (Knight 1994).

The formal approaches (shape grammars) and their contemporary counterparts (cellular automata and neural networks) (Coates and Langley 2007) are conditioned by the same theoretical framework of syntactic analysis of architectural configurations and variations of specific building types and historical styles (Ibrahim 2011). This literature is offered here as a brief introduction to how logic-analytical operations provided a framework to study architectural forms, based on the classification of variation by means of geometric representations of historic buildings, in order to provide the context for the following sections where the effort of the authors regarding the experimental use of neural networks for the annotations of variations of historic buildings is presented.

2 Semantic annotation methods in architecture heritage

ML (e.g., random forests) and the more recent deep learning (e.g., CNNs) methods have gained popularity, because of their successful application in a variety of tasks, with architecture analysis not being an exception. The exponential growth in computer vision, combined with the technological advances in 3D data documentation equipment - enabled the acquisition of high resolution and precise 3D data, e.g., point clouds and meshes (Georgopoulos and Ioannidis 2004), further boosting the interest in this area (Croce et al. 2020).

An early approach to building segmentation and classification was conducted in (Grilli et al. 2018), who unwrapped the UV map of a building's 3D model and used it for the training of a variety of decision tree-based methods. Parts of each UV map had to be manually annotated, signifying areas of interest (classes), followed by a preprocessing step for feature extraction. Once the decision trees were generated and trained on the selected features, they could be used for the classification of the remaining UV maps. This approach was tested on a small set of buildings, having satisfactory results in the recognition of structural elements and identification of degradation stages or restorations. The same method was later extended to operate on 3D data (coloured point clouds) too (Grilli and Remondino 2019). A key advantage of this method was the immediate visualisation of the results on the 3D model; on the other hand, a new decision tree needed to be built for each new test case.

Stathopoulou and Remondino (2019) used semantically segmented images to enhance feature selection during 3D reconstruction of buildings with photogrammetry, aiming to produce better models. Specifically, they created a dataset of historic building facade images, portraying a variety of architectural styles to ensure data diversity. After the manual annotation of the dataset (images), it was utilised for the training of a CNN, that in the end, would be able to semantically annotate new images. Morbidoni et al. (2020) exploited a variant of DGCNN (Wang et al. 2019) for the semantic annotation of partially annotated historical building point cloud scenes, acquired with Terrestrial Laser Scanning (TLS).

This work emphasises more on setting an easily reproducible pipeline that can be implemented with various DL methods, even when only a small number of buildings is available, rather than speed. The research proposes the fine-tuning of a pre-trained network to learn heritage-specific features in order to semantically segment the selected examples of Cypriot built heritage. By doing so, the research aims to introduce an alternative approach of structurally segmenting buildings, using a deep learning network.

2.1 A Convolutional Neural Network architecture for 3D processing of the geometric composition of built heritage

In this section, the selected network architecture of the proposed method, along with other design decisions made to increase its performance will be showcased. Before diving into the technical details though, the two datasets used for the training of the structure network will be presented, followed by the training procedure, and finally the performance evaluation metrics.

2.2 Datasets

A large amount of annotated data is required in order for a Deep Neural Network (DNN) to learn characteristic features for each structural component and at the same time generalise information (to avoid overfitting). This is not trivial, as the procedure of collecting and annotating this type of data is both tedious and laborious. Many activities have been carried out over the last few years but do not make their data publicly available to encourage future studies. An exception is the ArCH dataset (Matrone et al. 2020), which offers a small number of annotated model scenes, which however is not sufficient in size to train a DL method from scratch. Therefore, a different approach was considered by ANNFASS. This method consists of two stages - the pre-training and the fine-tuning - the first aiming to help the network learn generic structural features apparent in buildings, and the second to specialise on built heritage examples. Namely, the two acquired datasets, one for each stage, are the BuildingNet (Selvaraju et al. 2021) and ANNFASS (Deligiorgi et al. 2021).

BuildingNet consists of 2000 buildings of different architectural types (e.g., residential, religious) gathered from online sources and annotated with 31 unique part labels (e.g., wall, window). This dataset, being large enough in size, allows the training of a network for the building semantic segmentation task. The ANNFASS dataset is smaller in size and consists of historic buildings that were reality captured (3D documented) by means of structure-from-motion methods. The models were semantically labelled by cultural heritage experts in workshops organised in the context of the project through the use of a dedicated tool (fig. 1). Fig. 2 presents the initial set of labels, organised per historical architecture style, and table 1 shows the final list, where over-segmented categories were merged into broader ones. In the case of an

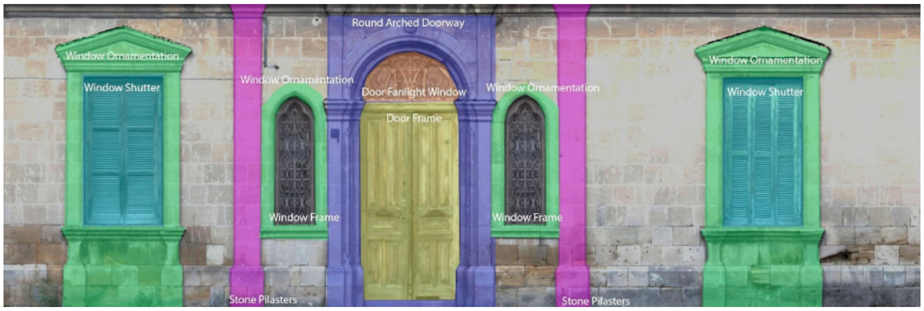
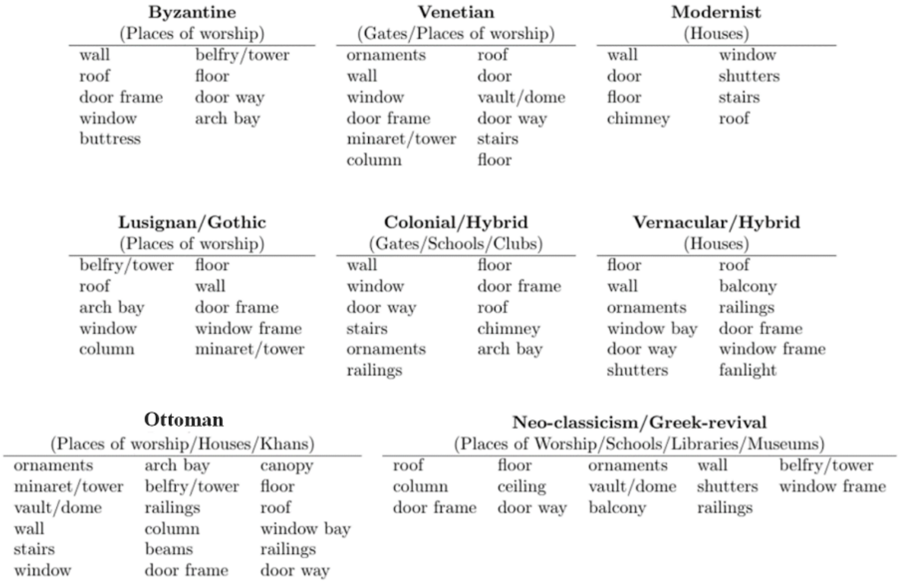


Figure 1. Example of Cypriot built heritage, semantically segmented and annotated by experts.

architectural component that could not be sufficiently characterised by the available labels, the component was assigned the label 0-undetermined.



Byzantine (Places of worship)	
wall	belfry/tower
roof	floor
door frame	door way
window	arch bay
buttress	

Venetian (Gates/Places of worship)	
ornaments	roof
wall	door
window	vault/dome
door frame	door way
minaret/tower	stairs
column	floor

Modernist (Houses)	
wall	window
door	shutters
floor	stairs
chimney	roof

Lusignan/Gothic (Places of worship)	
belfry/tower	floor
roof	wall
arch bay	door frame
window	window frame
column	minaret/tower

Colonial/Hybrid (Gates/Schools/Clubs)	
wall	floor
window	door frame
door way	roof
stairs	chimney
ornaments	arch bay
railings	

Vernacular/Hybrid (Houses)	
floor	roof
wall	balcony
ornaments	railings
window bay	door frame
door way	window frame
shutters	fanlight

Ottoman (Places of worship/Houses/Khans)		
ornaments	arch bay	canopy
minaret/tower	belfry/tower	floor
vault/dome	railings	roof
wall	column	window bay
stairs	beams	railings
window	door frame	door way

Neo-classicism/Greek-revival (Places of Worship/Schools/Libraries/Museums)				
roof	floor	ornaments	wall	belfry/tower
column	ceiling	vault/dome	shutters	window frame
door frame	door way	balcony	railings	

Figure 2. Initial list of architectural elements' labels that were used for the training and classification of architectural style of built heritage in Cyprus.

2.3 Network Training

The structural segmentation network's training happens in two stages, the pre-training and the fine-tuning. The network is initially trained with the BuildingNet dataset, learning to semantically segment building components. After this stage, the network is further trained on the ANNFASS dataset, to fine-tune its weights (layer connections) on the specific architecture features of the ANNFASS' built heritage. The authors'

Table 1. ANNFASS dataset's final list of labels

Architectural Component Labels			
1-wall	2-window frame	3-shutters	4-tower
5-door frame	6-column	7-railing	8-beam
9-floor	10-roof	11-stairs	12-ornamentation
13-dome	14-canopy	15-arch bay	16-buttress

decision to use a knowledge transferring technique (fine-tuning), instead of randomly initialising the network weights, permits the utilisation of a deep neural network (DNN) even when having very few instances. This way, rather than learning from scratch the monument features, the network only needs to slightly adjust the existing weights, to better suit the needs of the new dataset. This second stage of training is necessary, not only because the models contained in both datasets differ significantly (fig. 3), but also because of the variation in the ground truth labels (e.g. ornamentation that appears only in ANNFASS). Therefore, the network needs some additional training to learn features for the new labels and simultaneously calibrate the weights of the shared labels. Note that the weight fine-tuning process requires a small number of training epochs, in contrast to training from scratch.

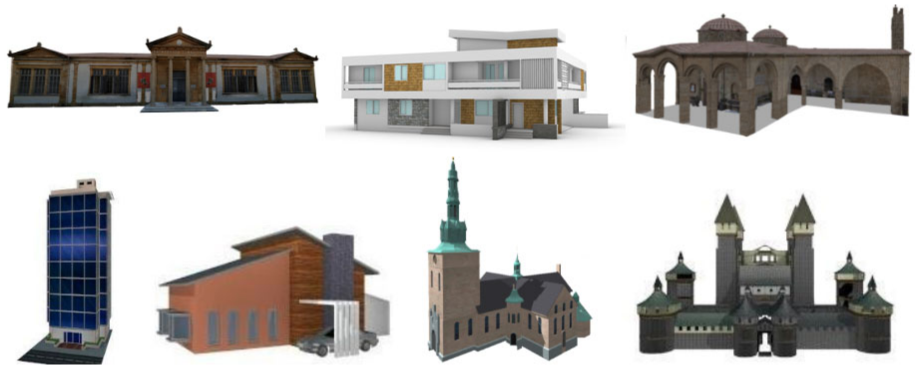


Figure 3. Example models from ANNFASS (top) and BuildingNet (bottom) datasets.

It is worth mentioning here that components with label undetermined are included in the input of the network, but are not considered during the loss and evaluation metric calculation. That is, the components are kept as part of the building, because by removing them from the geometry it would alter the network's perception of the inter-component relations, but they are masked out from the loss function and evaluation phase.

2.4 Evaluation metrics

The semantic segmentation task has occupied both the humanities and computer vision fields over the years, with a number of performance evaluation metrics being proposed. The presented research is using the overall accuracy and shape/part Intersection over Union (IoU) metrics.

Accuracy is used to measure the correctly predicted points over the entire test split's labelled points, while IoU is similar to F1-score - it is a combination of the network's precision and recall. The IoU for each part/label is computed for the test split monuments and averaged per label, resulting in the network's per part IoU, as presented in PartNet (Mo et al. 2018). The overall part IoU is the average of the per part IoUs. Shape IoU, on the other hand, is the average of IoUs computed per building. The former assesses the network's abilities to identify and differentiate the structural components, while the latter evaluates it on the monument level.

2.5 Octree-based High-Resolution Network (O-HRNet)

The work presented in this article regards a volumetric based network, specifically HRNet (Sun et al. 2019), as designed by Wang et al. (2021), using their memory and computation efficient octree implementation. Their input representation allows the processing of large input point clouds, by using high-resolution octrees, which reduce the chances of information loss (see fig. 4), unlike other networks, e.g., PointNet++ (Qi et al. 2017), or DGCNN (Wang et al. 2019), which apply subsampling to the input. Another benefit of using the octree implementation and convolutions introduced by (Wang et al. 2017), is the freedom to add more features as input, e.g., colour, in contrast with other volumetric methods which solely rely on the normal vectors. To sum up, in this work an Octree-based HRNet (O-HRNet) is employed to semantically annotate building components, using as input their octree representation.

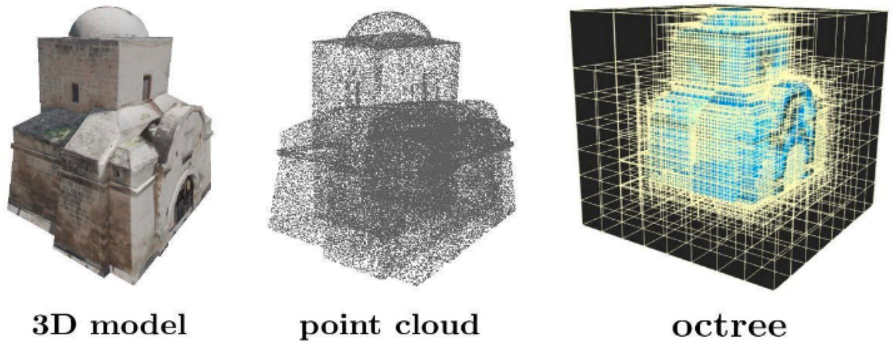


Figure 4. Representation of a monument's 3D model (left) using a point cloud (middle) and an octree (right). Octrees prove to be an equally good approximation of the input, as point clouds.

An important aspect of the O-HRNet's architecture is the parallel multi-resolution feature extraction and inter-resolution feature exchange (fig. 5). The network processes different resolutions of the input model at the same time and shares information among the different layers, by exchanging the extracted features, to help improve the network's localisation and shape perception. The multi-resolution handling is achieved with the use of branches (shown with green, lilac and yellow colours in fig. 5), and the feature exchange through up/downsampling links (red/blue arrows) between the branches. By implementing these two simple concepts (multi-resolution and information sharing) within the network, it benefits from the infusion of information learned in different resolutions, using them to get a better understanding of the input and extract more meaningful, component/part specific features.

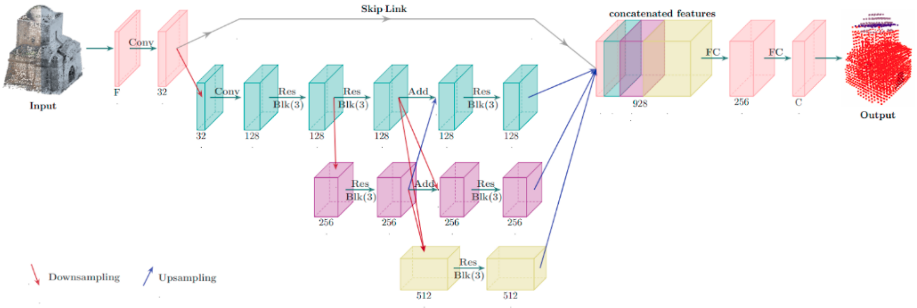


Figure 5. O-HRNet's architecture, starting from the input point cloud with/without colour features, which is internally transformed into an octree, to be processed by the network's layers. Then, the octree is simultaneously processed at three consecutive branches of different resolutions to extract both global and local features that will be passed through two FC layers to produce part labels for the input.

Once the processing of the features is completed, through a sequence of convolutions and residual blocks, features from all resolutions are concatenated and passed through two Fully Connected (FC) layers to generate labels for each leaf octant. Then, a post-processing step is taken to interpolate the octant labels back to the initial point cloud. Another important aspect of the presented architecture is the use of a weighted loss function. Buildings, by design, are an imbalanced dataset, since some components appear more frequently or cover larger areas, e.g., wall vs ornamentation. A popular way of handling these cases is the introduction of a label-dependent weighting factor in the loss function, which is the approach followed here. These weights are inversely proportional to a label's frequency and serve as a counterbalance for the rarity of some labels. In other words, mislabelling a rarer label is more 'costly' than that of a more common one, making the network pay equal attention to all labels, in order to reduce the training error.

2.6 Results

The previous sections explained the proposed pipeline, along with the network (O-HRNet) the authors applied to this pipeline, to identify the structural elements of the selected buildings. Below, a series of experiments is presented to show (a) that fine-tuning is better than training from scratch and (b) that the proposed pipeline can be used with other neural networks.

Looking at tables 2 and 3, which contain the overall and per label evaluation of the experiments, respectively, it becomes very clear that fine-tuning (independent of the network) outperforms scratch training. This is due to the dataset's small size, which prevents us from training the network for too long, as it would overfit. It is obvious that the network in the small training time that is given manages to learn features only for components that appear frequently (e.g., wall) or that have characteristic geometry (eg., tower), hence for the very low Part IoU. Therefore, this proves that statement (a) is correct, i.e. using a learning migration technique helps the network learn more specific features, even when the dataset is small.

Regarding the second statement made above, we experimented with substituting O-HRNet with another popular network, PointNet++, to prove that the pipeline is agile and can be applied to any NN. It is worth mentioning that we quadrupled the resolution of PointNet++, by taking more representative points and reducing the neighbourhood radii by 4 so that we achieve a higher resolution of the input and manage to capture detail of the smaller components (e.g., ornamentation).

Indeed PointNet++ can be used in the proposed pipeline, and has better performance than the trained from scratch O-HRNet. The large gap between the 2 fine-tuned models is highly correlated with the network's input. Both networks accept point clouds of the same size as input and use a multi-resolution technique, though PointNet++ performs a subsampling step prior to the feature extraction making it more prone to information (component) loss. Therefore, O-HRNet performs better because it manages to capture more details of the input model, but any other network can be used in its place.

Table 2. Comparison of evaluation metrics for the different experiments carried in this article using PointNet++ or O-HRNet. F refers to the use of fine-tuning and in bold are the best performing scores.

Method	Accuracy	Shape IOU	Part IoU
O-HRNet	57.7	13.3	6.6
O-HRNet & F	**68.5**	**24.8**	**28.0**
PointNet++ & F	58.4	14.9	17.2

Table 3. Per label IoUs for the different experiments carried in this article using PointNet++ or O-HRNet. F refers to the use of fine-tuning. Column numbers correspond to the label ids from table 1 and in bold are the highest part IoUs.

Method/Label	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
O-HRNet	54.8	0.0	0.0	2.6	0.0	0.0	0.0	0.6	2.1	45.8	0.0	0.0	0.0	0.0	0.0	0.0
O-HRNet & F	**63.8**	**40.2**	7.1	**64.0**	**21.4**	19.2	**26.2**	6.0	**51.9**	**64.1**	**8.9**	0.0	**32.4**	0.0	**36.6**	**5.8**
PointNet++ & F	56.0	15.1	**11.9**	3.3	4.9	**21.3**	10.5	**12.6**	36.7	48.3	1.8	**4.9**	30.2	0.0	18.6	0.0

To conclude, an agile pipeline was presented that is easy to reproduce, uses basic features (not related to the capturing method), has good performance in structural segmenting monuments and most importantly does not need retraining for every new instance that we want to segment. It was shown that different networks can be applied in the pipeline, specifically we tested PointNet++ and O-HRNet. Though, the latter has a significantly better performance than the former, because of its high-resolution representation of the input.

3 Style annotation methods in built heritage

With the advances of ML, efforts were made of using said methods in building style classification. Early works have used Computer Vision (CV) methods in order to extract useful features, like the Harris-Laplacian corner detector, the Histogram of Oriented Gradients (HOG) and Fisher Vectors, that when combined with ML techniques can classify a small subset of buildings (Montoya Obeso et al. 2016; Shalunts 2015; Mathias et al. 2012). More recent works applied DL - precisely CNNs - directly on building images, in order for example to help the real estate industry to taxonomize the buildings, to estimate their values (Xia et al. 2020) or the building's age (Zeppelzauer et al. 2018). Another recent study on cultural heritage buildings uses DL to classify the building's stones in order to identify the structural evolution of the buildings (Mesanza-Moraza et al. 2020).

It is important to stress out the limitation of available data. The aforementioned works made great efforts to use DL and classify data in 15-35 categories, rather than 3 categories as in earlier works, however the data used are still in the range of 300-500 images. This amount of data limits the complexity of the DL models that can be used. Additionally there is absence of research on 3D data in the building heritage domain. However there is some work done in the CV community that can classify 3D objects, including a set of 329 buildings into their style. The proposed state-of-the-art method on that data creates multi-view representations from multiple patches of the objects, extracts their HOG features and uses Partially Shared Latent Factor (PSLF) learning (Yu et al. 2018).

3.1 Datasets

Instead of using 'typical' examples of the high period of each architectural style to both train the NN and test their performance, this research opted for an analysis of important monuments of Cypriot heritage which exhibit a mélange (Given 2005) - a mixture of types of building components classified in various architecture styles. However, as in some other works, the classification here happens on the building's component level. The reason for this decision is twofold; first, hybrid buildings are no longer a problem for the method to classify, and second, the amount of data is increased if components are considered rather than buildings.

As with the pipeline in Section 2, the BuildingNet dataset was utilized to train a DNN. Only components that were indicative for stylistic influences, and which were straightforward to segment, were used in the experiments, namely columns, doors, windows, domes and towers. In order to test the trained network annotated data had to be collected. The buildings from Deligiorgi et al. (2021), along with 100 buildings from the BuildingNet collection were annotated by experts. The final data used are presented in table 4.

The data are split into 3 sets. Set A represents all the unlabeled data (coming from 1840 BuildingNet buildings). This set can be used within the context of Knowledge Transfer, i.e. in a pretext task (a task different from the ultimate one). The rest of the data, which are labeled, are split into set B and C, for training and testing, respectively, the style classifier (the downstream task). Before splitting the labeled dataset, the following considerations were made:

- Components from the same building should not appear in different sets.
- Both training and testing should have data from all styles.

- The dataset is not balanced in any way, and ideally the training split should consist of more buildings and more building elements than the test split.
- Since the labeled dataset is small, cross-validation must be used.

Table 4. Final list of style labels found in models of ANNFASS and subset of BuildingNet datasets

		Labels on both sets					
Style	#buildings	Unique components					
		#column	#dome	#door	#tower	#window	total
1-baroque	7	1	6	1	7	3	18
2-byzantine	5	3	4	2	6	3	18
3-colonial	10	0	0	9	3	12	24
4-gothic	12	5	0	5	9	14	33
modernist	3	0	0	3	0	3	6
5-neoclassicism	9	8	2	6	4	10	30
6-ottoman	19	8	13	12	17	14	64
pagoda	3	2	0	1	3	1	7
renaissance	2	1	0	1	4	1	7
7-romanesque	7	1	2	4	11	6	24
russian	3	0	4	1	6	2	13
venetian	12	1	0	4	0	2	7
unlabeled	1480	389	178	996	425	1402	3390
Total within 7 classes	69	26	27	39	57	62	211

3.2 Method

The fine-tuning approach followed for the structure network (Section 2) could not be used for style annotation, since BuildingNet does not contain any style labels. Therefore, another form of Knowledge Transfer is used, specifically, a DNN is trained on the task of shape reconstruction, in order to achieve dimensionality reduction. The autoencoder network architecture is used to convert the input data into an intermediate (latent) representation of lower dimension, that can be employed to re-generate the input. This architecture creates a bottleneck that captures only the most useful features needed to reconstruct the data (see fig. 6).

In this approach, sparse representations of point clouds are used and being processed with sparse convolutions (Minkowski Networks introduced in Choy et al. (2019)). This type of network achieves great performance in many 3D tasks. Unlike dense tensors when using sparse tensors, the number of features can vary per input/output. This makes it possible for point clouds with different sizes to be processed by the same network (unlike, for example, in PointNet). This is desirable, since some shapes are more detailed than others, and more points are needed to represent them than other, simpler shapes.

Sparsity regards using the whole space and time capacity for meaningful information, in order words, keeping only the locations of a 3D model that do include some data. Therefore, the input model in the form of an arbitrary sized point cloud, is quantized (given some unit length) into cubes. Each unit cube is then represented

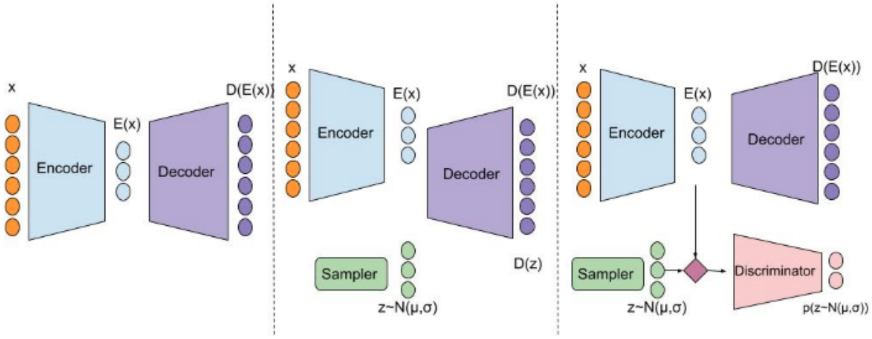


Figure 6. Left: Autoencoder Architecture. Middle: Variational Autoencoder. During the training phase, the encoder is used; at generation phase, the sampler is used. Right: Adversarial Autoencoder. The Discriminator classifies whether the latent vector comes from Sampling or Real Data distribution. Both VAE and AAE force the latent representation to belong in a well organized space.

by its average features, and unique label (the unit cube must be small so as cubes with multiple labels do not often occur; in other cases the cube is rejected). The sparse tensor can now hold only the cubes with non-empty data, and represent them by their mean coordinate. When a sparse tensor is processed by a sparse convolutional layer, the output is a coarser 3D model and when it is processed by a sparse transpose convolutional layer, the output is a more detailed 3D model.

The steps of the method presented are:

- First train a Minkowski-based ResNet34 Variational Autoencoder using dense point clouds of the 3D components belonging in set A.
- Then pass the labeled components (set B,C) through the trained network and extract their latent codes (features).
- Finally, use the latent codes of components in set B to train a Support Vector Machine on style classification, and test its performance on set C.

This pipeline can be seen in fig. 7.

3.3 Evaluation

To evaluate the proposed method, Recall and Precision are used, along with their harmonic mean, the F1 score. Recall indicates the fraction of correctly predicted components for a class, over all components predicted in that class. Precision indicates the fraction of correctly predicted components for a class, over all components that truly belong in that class.

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

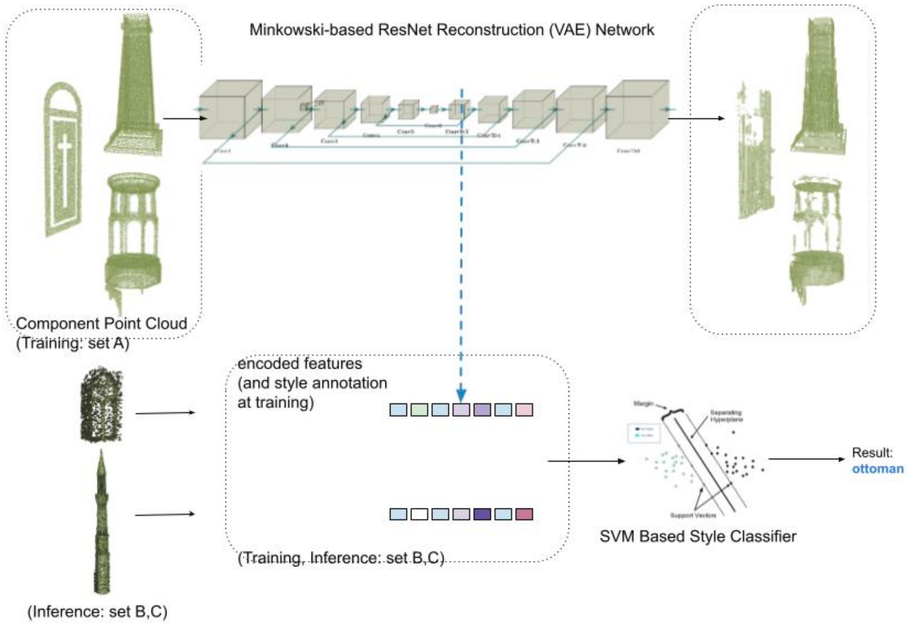


Figure 7. The NN architecture of the reconstruction pretext task trained on set A is shown on top. The point clouds on the right are real outputs of the NN. The results are noisy especially for the windows and doors that appear often as plain meshes or planes with different drawings on them. The pipeline of extracting the component's feature vector from the latent representation of the VAE, and using set B and C for training/inference of the SVM is shown on the bottom.

$$F1Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

Since the annotated dataset is small, 5-fold cross validation is used, and the experiments are repeated 5 times. The method is compared to the conventional state-of-the-art method of Yu et al. (2018). Three variants of the method are also tested; one where the Minkowski-based autoencoder is replaced by a PointNet-based autoencoder, and two where the pretext task of shape reconstruction is replaced by the structure segmentation task, either with a Minkowski-based HRNet or an Octree-based HRNet (adopted from Section 2).

3.4 Results

The results (see tables 5 and 6) indicate that the proposed method achieves the best performance. All DL approaches on 3D shapes outperform the method of Yu et al. (2018) which is based on conventional Computer Vision and Machine Learning. Both Minkowski based networks achieve better performance than the PointNet based reconstruction pretext task, and the Minkowski based semantic segmentation network

Table 5. Precision (P) and Recall (R) levels per class and over all classes for each method. The results are calculated over 5 repetitions of 5-fold cross validation.

		Baroque	Byzantine	Colonial	Gothic	Neo classicism	Ottoman	Romanesque	Mean
Average Test Components		6	7	12	18	12	16	9	-
Yu et al. (2018)	P	0.140	0.048	0.345	0.24	0.204	0.084	0	0.152
	R	0.298	0.069	0.294	0.141	0.089	0.145	-	0.148
PointNet AAE	P	0.089	0.068	0.163	0.24	0.162	0.166	0.152	0.148
	R	**0.136**	0.112	0.148	0.15	0.16	0.137	0.16	0.143
Minkowski VAE	P	**0.112**	0.007	**0.542**	**0.542**	**0.218**	0.22	0.189	**0.261**
	R	0.107	0.005	0.421	**0.421**	**0.217**	**0.587**	0.151	**0.273**
Minkowski PS	P	0	0	0.162	0.22	0.141	0.197	**0.87**	0.227
	R	0	0	0.183	0.066	0.19	0.431	**0.3**	0.167
Octree PS	P	-	**0.245**	0.241	0.186	0.059	**0.259**	0.209	0.199
	R	0	**0.234**	**0.698**	0.076	0.043	0.213	0.252	0.216

Table 6. Average and maximum F1 score obtained over the 5 repetitions for each method, along with the latent dimension used for component representations by the style classifier. Note that the average F1 score obtained by a random assignment classification is 0.078.

Method	Yu et al. (2018)	3D PointNet AAE (Reconstruction)	3D Minkowski VAE (Reconstruction)	3D Minkowski HRNet (Structure Segmentation)	3D Octree HRNet (Structure Segmentation)
mean F1	0.122	0.134	**0.222**	0.164	0.173
max F1	0.147	0.169	**0.237**	0.183	0.188
Latent Dimension	900	2048	1024	64	928

achieves similar performance to its Octree-based counterpart which uses bigger latent dimension. This indicates - as expected - that Minkowski based networks are more powerful than their counterparts. The results suggest that the task of shape reconstruction is better suited for the style analysis, rather than the task of semantic segmentation; however, both achieve good performance, and a method that utilizes both, would probably be the best approach.

Similarly to the structure analysis pipeline that was presented in the article, this pipeline can be used by different types of DNNs, and additionally, both reconstruction and part segmentation pretext task can be utilized. Note that in the style classification experiments conducted by the research, point clouds with no color information were used. It is expected that adding the colour might improve the performance of the networks. Further improvements that might improve performance include 1) the pretext task of whole building reconstruction, and 2) the integration of chronological periods in style labels when training the ML models (since styles are correlated to each other, this could lead to better style disambiguation (Yi et al. 2020)). Note that the performance of style classification presented is low, but these results are close to the findings in Yi et al. (2020), where a comparable dataset (similar amount of classes, training and test samples) was used. This indicates how challenging the task of architectural style classification is.

4 Conclusion and Future Work

ML methods provided new tools to researchers for the segmentation and semantic annotation of architectural heritage which under the right conditions could contribute to classification processes. The advantage of the presented methods, which are based on Knowledge Transfer in DNNs, is their speed in training the final model and the avoidance of the time consuming preprocessing steps of feature analysis and extraction, which are required by the majority of ML methods. More importantly, these methods are generic and agile enough, allowing their operation with different architectural styles/components, avoiding the common need for generation and training of a new model for every new heritage building that is added to the dataset.

Concluding, it is worth mentioning that the ability to reason about a building's form in a virtual environment is crucial for efficient documentation and cataloguing, education, as well as for facilitating remote study in the humanities. The ANNFASS project brings 3D CNNs and the annotated building dataset together through an online platform[3] that will enable scholars, researchers and students to remotely access 3D digital repositories that are dynamically enriched with additional (new) 3D data in the future.

Notes

1. http://annfass.cs.ucy.ac.cy
2. https://www.research.org.cy/en/
3. https://annfass-srv.cs.ucy.ac.cy/home

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 739578 and under grant agreement No EXCELLENCE/1216/0352 and the Government of the Republic of Cyprus through the Directorate-General for European Programmes, Coordination and Development. The work was also partially funded by Adobe, and project TIN2017-88515-C2-2-R from Ministerio de Ciencia, Innovation y Universidades, Spain.

References

Jiménez-Badillo, Ruíz-Correa, and García-Alfaro. Developing a recognition system for the retrieval of archaeological 3d models. In *Fusion of Cultures: Proceedings of the 38th. Annual Conference on Computer Applications and Quantitative Methods in Archaeology, Granada, Spain*, pages 325–331, 2010.

Laura Baratin, Sara Bertozzi, Elvio Moretti, and Michele Spinella. Gis and 3d models as support to documentation and planning of the baku historical centre (republic of azerbaijan). *International Journal of Heritage in the Digital Era*, 1:71–76, 1 2012. ISSN 2047–4970. doi: 10.1260/2047-4970.1.0.71. URL http://journals.sagepub.com/doi/10.1260/2047-4970.1.0.71.

W.J. Mitchell, R.S. Liggett, S.N. Pollalis, and M. Tan. Integrating shape grammars and design analysis. In *CAAD Futures '91 Conference Proceedings*, pages 17–32, 1991. ISBN 3-528-08821-4.

J.S. Gero. *Design Computing and Cognition'20*. Springer International Publishing, 2020. URL https://link.springer.com/book/9783030906245.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

Y. Yoshimura, B. Cai, Z. Wang, and C. Ratti. Deep learning architect: Classification for architectural design through the eye of artificial intelligence. *arXiv*, 2018. doi: 10.1007\/978-3-030-19424-6_14. URL https://www.researchgate.net/publication/332997690_Deep_Learning_Architect_Classification_for_Architectural_Design_Through_the_Eye_of_Artificial_Intelligence.

J Duarte and J Rocha. A grammar for the patio houses of the medina of marrakech - towards a tool for housing design in islamic contexts, 2006.

Olivier Teboul, Iasonas Kokkinos, Loic Simon, Panagiotis Koutsourakis, and Nikos Paragios. Shape grammar parsing via reinforcement learning. In *CVPR 2011*, pages 2273–2280, 2011. doi: 10.1109/CVPR.2011.5995319.

Jose Llamas, Pedro M. Lerones, Eduardo Zalama, and Jaime Gómez-García-Bermejo. Applying deep learning techniques to cultural heritage images within the inception project. In *EuroMed*, volume 10059 LNCS, pages 25–32, 2016. doi: 10.1007/978-3-319-48974-2_4.

Gayane Shalunts. Architectural style classification of building facade towers. *Advances in Visual Computing*, pages 285–294, 2015. doi: 10.1007/978-3-319-27857-5_26. URL https://link.springer.com/chapter/10.1007/978-3-319-27857-5_26#:~:text=Towers%20are%20architectural%20structural%20elements,geographically%20widely%20spread%20in%20Europe.

Lionel March and Philip Steadman. *Geometry and Environment: an introduction to spatial organization in design.* Routledge, 2021. URL https://www.routledge.com/The-Geometry-of-Environment-An-Introduction-to-Spatial-Organization-in/March-Steadman/p/book/9780367360245.

F. Chiabrando, M. Lo Turco, and F. Rinaudo. Modeling the decay in an hbim starting from 3d point clouds. a followed approach for cultural heritage knowledge. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W5:605–612, 08 2017. doi: 10.5194/isprs-archives-xlii-2-w5-605-2017. URL https://ui.adsabs.harvard.edu/abs/2017ISPAr62W5..605C/abstract.

Hélène Macher, Tania Landes, and Pierre Grussenmeyer. From point clouds to building information models: 3d semi-automatic reconstruction of indoors of existing buildings. *Applied Sciences*, 7:1030, 10 2017. doi: 10.3390/app7101030. URL https://www.mdpi.com/2076-3417/7/10/1030.

Chao Wang, Yong K. Cho, and Changwan Kim. Automatic bim component extraction from point clouds of existing buildings for sustainability applications. *Automation in Construction*, 56:1–13, 08 2015. doi: 10.1016/j.autcon.2015.04.001. URL https://www.sciencedirect.com/science/article/pii/S0926580515000734.

Michael Given. *Architectural styles and ethnic identity in medieval to modern Cyprus*, chapter 25, pages 207–213. Oxbow Books, 01 2005. ISBN 1842171682.

M. Given and J.S. Smith. *The Sydney Cyprus Survey Project: Social Approaches to Regional Archaeological Survey*. Cotsen Institute of Archaeology, University of California, 2003. ISBN 9781931745048. doi: 10.2307\/4150093.

K.W. Schaar, Michael Given, and G. Theocharous. *Under the clock: colonial architecture and history in Cyprus, 1878-1960. Nicosia: Bank of Cyprus.* Bank of Cyprus, 01 1995.

Yannis Hamilakis. Through the looking glass: nationalism, archaeology and the politics of identity. *Antiquity*, 70:975–978, 12 1996. doi: 10.1017/s0003598x00084271. URL https://www.cambridge.org/core/journals/antiquity/article/abs/through-the-looking-glass-nationalism-archaeology-and-the-politics-of-identity/6AFEA80895E39372352C1BB5DDA885BA.

George Nicholas Stiny. Computing with form and meaning in architecture. *Journal of Architectural Education*, 39(1):7–19, 1985. doi: 10.1080/10464883.1985.10758382. URL https://doi.org/10.1080/10464883.1985.10758382.

Terry W. Knight. *Transformations in Design: A Formal Approach to Stylistic Change and Innovation in the Visual Arts*, volume New York: Springer-Verlag. Cambridge University Press, 1994.

Paul S. Coates and Phillip Langley. Meta-cognitve mappings, growing neural networks for generative urbanism. *Generative Art Conference*, 2007.

Mohamed Ibrahim. *Structuring the design studio education Crafting the projects of the beginning studio using shape grammars*. PhD thesis, Alexandria University, 07 2011.

Andreas Georgopoulos and Charalambos Ioannidis. Photogrammetric and surveying methods for the geometric recording of archaeological monuments photogrammetric and surveying methods for the geometric recording of archaeological monuments. In *International Federation of Surveyors (FIG)*, pages 22–27, 2004.

V. Croce, G. Caroti, L. De Luca, A. Piemonte, and P. Véron. Semantic annotations on heritage models: 2d/3d approaches and future research challenges. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B2-2020:829–836, 2020. doi: 10.5194/isprs-archives-XLIII-B2-2020-829-2020. URL https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLIII-B2-2020/829/2020/.

E. Grilli, D. Dininno, G. Petrucci, and F. Remondino. From 2d to 3d supervised segmentation and classification for cultural heritage applications. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2:399–406, 05 2018. doi: 10.5194/isprs-archives-xlii-2-399-2018. URL https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLII-2/399/2018/.

Eleonora Grilli and Fabio Remondino. Classification of 3d digital heritage. *Remote Sensing*, 11(7), 2019. ISSN 2072-4292. doi: 10.3390/rs11070847. URL https://www.mdpi.com/2072-4292/11/7/847.

E.-K. Stathopoulou and F. Remondino. Semantic photogrammetry – boosting image-based 3d reconstruction with semantic labeling. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W9:685–690, 2019. doi: 10.5194/isprs-archives-XLII-2-W9-685-2019. URL https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLII-2-W9/685/2019/.

C. Morbidoni, R. Pierdicca, R. Quattrini, and E. Frontoni. Graph cnn with radius distance for semantic segmentation of historical buildings tls point clouds. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIV-4/W1-2020:95–102, 2020. doi: 10.5194/isprs-archives-XLIV-4-W1-2020-95-2020. URL https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.

net/XLIV-4-W1-2020/95/2020/.

Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph cnn for learning on point clouds, 2019.

F. Matrone, E. Grilli, M. Martini, M. Paolanti, R. Pierdicca, and F. Remondino. Comparing machine and deep learning methods for large 3d heritage semantic segmentation. *ISPRS International Journal of Geo-Information*, 9(9), 2020. ISSN 2220-9964. doi: 10.3390/ijgi9090535. URL https://www.mdpi.com/2220-9964/9/9/535.

Pratheba Selvaraju, Mohamed Nabail, Marios Loizou, Maria Maslioukova, Melinos Averkiou, Andreas Andreou, Siddhartha Chaudhuri, and Evangelos Kalogerakis. Buildingnet: Learning to label 3d buildings. *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Marissia Deligiorgi, Maria I. Maslioukova, Melinos Averkiou, Andreas C. Andreou, Pratheba Selvaraju, Evangelos Kalogerakis, Gustavo Patow, Yiorgos Chrysanthou, and George Artopoulos. A 3d digitisation workflow for architecture-specific annotation of built heritage. *Journal of Archaeological Science: Reports*, 37:102787, 2021. ISSN 2352-409X. doi: https://doi.org/10.1016/j.jasrep.2020.102787. URL https://www.sciencedirect.com/science/article/pii/S2352409X20305782.

Kaichun Mo, Shilin Zhu, Angel X. Chang, Li Yi, Subarna Tripathi, Leonidas J. Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding, 2018.

Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation, 2019.

Peng-Shuai Wang, Yu-Qi Yang, Qian-Fang Zou, Zhirong Wu, Yang Liu, and Xin Tong. Unsupervised 3d learning for shape analysis via multiresolution instance discrimination, 2021.

Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space, 2017.

Peng-Shuai Wang, Yang Liu, Yu-Xiao Guo, Chun-Yu Sun, and Xin Tong. O-cnn: octree-based convolutional neural networks for 3d shape analysis. *ACM Transactions on Graphics*, 36(4):1–11, Jul 2017. ISSN 1557-7368. doi: 10.1145/3072959.3073608. URL http://dx.doi.org/10.1145/3072959.3073608.

Abraham Montoya Obeso, Jenny Benois-Pineau, Alejandro Ramirez-Acosta, and Mireya Vázquez. Architectural style classification of mexican historical buildings using deep convolutional neural networks and sparse features. *Journal of Electronic Imaging*, 26: 011016, 12 2016. doi: 10.1117/1.JEI.26.1.011016.

Markus Mathias, Andelo Martinovic, Julien Weissenberg, S. Haegler, and Luc Van Gool. Automatic architectural style recognition. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XXXVIII-5/W16:171–176, 09 2012. doi: 10.5194/isprsarchives-XXXVIII-5-W16-171-2011.

Bing Xia, Xin Li, Hui Shi, Sichong Chen, and Jiamei Chen. Style classification and prediction of residential buildings based on machine learning. *Journal of Asian Architecture and Building Engineering*, 19(6):714–730, 2020. doi: 10.1080/13467581.2020.1779728. URL https://doi.org/10.1080/13467581.2020.1779728.

Matthias Zeppelzauer, Miroslav Despotovic, Muntaha Sakeena, David Koch, and Mario Döller. Automatic prediction of building age from photographs. *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, Jun 2018. doi: 10.1145/3206025.3206060. URL http://dx.doi.org/10.1145/3206025.3206060.

Amaia Mesanza-Moraza, Ismael García-Gómez, and Agustín Azkarate. Machine learning for the built heritage archaeological study. *Journal on Computing and Cultural Heritage*, 14: 1–21, 12 2020. doi: 10.1145/3422993.

Fenggen Yu, Yan Zhang, Kai Xu, Ali Mahdavi-Amiri, and Hao Zhang. Semi-supervised co-analysis of 3d shape styles from projected lines. *ACM Transactions on Graphics*, 37(2): 1–17, Jul 2018. ISSN 1557-7368. doi: 10.1145/3182158. URL http://dx.doi.org/10.1145/3182158.

Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019.

Y.K. Yi, Y. Zhang, and J. Myung. House style recognition using deep convolutional neural network. *Automation in Construction*, 118, 2020.