Review

Advancements in AI-Driven 3D Reality Capture: Transforming Architectural Digitisation and Modelling

Kai Zhang and Francesco Fassi *

Posted Date: 18 June 2024

doi: 10.20944/preprints202406.1162.v1

Keywords: digitalization; artificial intelligence; 3D modelling; object detection; semantic segmentation; machine learning; deep learning

 doi:10.20944/preprints202406.1162.v1

Review

Advancements in AI-Driven 3D Reality Capture: Transforming Architectural Digitisation and Modelling

Kai Zhang and Francesco Fassi *

3D Survey Group, ABC Department, Politecnico di Milano, Via Ponzio 31, 20133 Milano, Italy; kai.zhang@polimi.it (K.Z.); francesco.fassi@polimi.it (F.F.)

* Correspondence: francesco.fassi@polimi.it

Abstract: 3D reality capturing has demonstrated increased efficiency and consistently accurate outcomes in architectural digitisation. Nevertheless, despite advancements in data collecting, 3D reality capturing still lacks full automation, especially in the post-processing and modelling phase. Artificial intelligence (AI) has been a significant focus, especially in computer vision, and tasks such as image classification and object recognition might be beneficial to the digitisation process and its subsequent utilisation. This article aims to examine the potential outcomes of integrating AI technology into the field of 3D reality-capturing, with a particular focus on its use in architectural scenarios. The main methods used for data collection are laser scanning (static or mobile) and photogrammetry. As a result, image data including RGB-D data (files containing both RGB colours and Depth information) and point clouds have become the most common raw datasets available for object mapping. The study comprehensively analyses the current use of 2D and 3D deep learning techniques in documentation tasks, particularly downstream applications. It also highlights the ongoing research efforts in developing real-time applications with the ultimate objective of achieving generalisation and improved accuracy.

Keywords: digitalization; artificial intelligence; 3D modelling; object detection; semantic segmentation; machine learning; deep learning

1. Introduction

An intelligent mapping system is urgently required because, although the rapid development of reality capture technology has stimulated documentation activities in recent years, the data processing still requires extensive human intervention.

In-situ investigations are expensive in terms of labour cost and time, although now, many technologies offer fast, automatic, and complete digitalisation capabilities. The elaboration works, including data processing, 3D reconstruction, registering, and further downstream works like information modelling and quality control for reality capturing, are mainly done off-site and are, till today, time-consuming because they are manual. Therefore, the urge to improve the productivity of 3D modelling leads to two issues that require immediate attention: instant (real-time) investigation feedback and the interpretation of interested objects in 3D.

AI technologies are witnessed to be capable of automating and speed up many time-consuming human labour processes. For example, the segmentation and labelling of semantic meanings might be the most relevant and mature tools nowadays usable to help the process of digitalised 3D spatial data.

Specifically, for 2D images, mature deep learning (DL) models (e.g. YOLOv10 [1]) can run at 200 frames per second (FPS) or more, achieving real-time object detection. For 3D point clouds, machine learning models achieved 90%+ classification accuracy based on large ranges' geometric features (e.g. DGCNN [2]). Much of the research in the last few years has been seen integrating the 2D and 3D

 doi:10.20944/preprints202406.1162.v1

automatic processing tools, mapping objects and items in the spaces, and giving corresponding semantic information, making possible quick assessment of the investigation site alongside the data acquisition activities.

These AI applications can greatly benefit 3D reconstruction activities and further data harvesting. In fact, 3D data with semantic meanings facilitate viewing and editing digital assets, enabling comparison and integration with previous surveys or the BIM model, and contributing to the BIM population and fieldwork monitoring. Considering the characteristics of different data types and corresponding technologies, combining artificial intelligence processing with reality capturing can help with real-time monitoring of building construction, facility control, and activities in many other scenarios.

This paper collects and organises noteworthy research relevant to the interests of the reality-capturing field in architectural scenarios. The related fields have presented many promising tools, but they are mostly constrained to specific scenarios. Listing the techniques and applications reveals some common approaches and a trend of research towards deep articulation can be found.

1.1. An Initial Overview of AI Techniques for Digitalisation

AI, especially machine learning, usually refers to the field of research in computer science that enables machines to learn from given data and generate responses to defined demands.

Machine learning is closely related to statistical models. It fits input data and generates classification (predicting predefined class labels), segmentation (dividing an image or data into multiple segments or regions), or regression (predicting continuous numerical values) strategies. Some widely used algorithms are Decision Trees, Random Forests (RF) [9,24], Support Vector Machines (SVM) [25], K nearest neighbour (KNN), Naïve Bayes (NB), etc. Generally, these statistical models are simple and easily explainable. However, these models must be expanded and integrated with sophisticated feature extractors to accurately predict more complex inputs, like images and point clouds. These feature extractors induce context information to the data points (e.g. pixels in images) but at this stage were mostly human engineered.

Deep learning, as a subsector of machine learning, largely increased the model complexity and further automatised the learning process, allowing models to handle intricate data efficiently. Nowadays, DL tools can be expected to perform more complicated tasks like object detection (identifying and localising objects within images), pose estimation (determining the positions of subject's key points, e.g., joints of the human body), text generation (producing sequences of words to create contextually relevant text), etc.

DL models learn the context information automatically. Initially, based on the data type, different approaches were developed, therefore two main research orientation: computer vision (CV) processes 2D data, Natural Language Processing (NLP), which processes text data. Concretely, considering the task and the data being analysed, inputs can be classified based on their dimensionality, including:

(1) One-dimensional inputs: Sensor data, including data from sensors such as accelerometers, gyroscopes, and temperature sensors; audio data, including speech, music, and other audio recordings; and some time series data.
(2) Two-dimensional inputs: Image data, including photographs, drawings, spectrograms, etc. Image data can be analysed using CV techniques, which involve processing and analysing visual data. The sequence of image data or video can be considered multi-dimensional.
(3) Three-dimensional inputs: Point cloud data and 3D scans of objects or environments, such as buildings, landscapes, and industrial equipment. Point cloud data can be analysed using algorithms that are specifically designed for 3D point cloud data. Addressed in the work of PointNet [3], points from the Euclidean space are unordered, which greatly differentiates from pixel arrays in 2D images or voxel arrays in volumetric grids. Changes of data feeding order will not change the point cloud essence. Another popular representative format of 3D is RGB-D data, which combines RGB colour data with depth data (D). Depending on the methods (either structure light or time-of-flight), the camera acquires the precise distance of the object's surfaces from the specific viewpoint where RGB information is collected.

 doi:10.20944/preprints202406.1162.v1

(4) Multi-dimensional inputs: Text data, including articles, reviews and more, could have many features such as word frequency, word length, and syntactic structure. Text data can be analysed using NLP techniques, which involve processing and analysing natural language data.

Generally, different types of inputs require specific data-analysing tools and feature extractors. In some cases, the AI methods can also generalize well to other types of inputs, e.g., in the NLP field the "attention" mechanism, which uses "transformer" modules to calculate attention weights of word within specific section of the sentence, was applied in image object detection [4] and the convolution filter application in the 3D point cloud [3,5].

With the growth of computational power, AI methods like DL have made data processing like classification able to achieve satisfying results with less labour and time. Since introducing by Lecun et al., 1989 [6] the Convolutional Neuron Network (CNN), i.e., a regularised version of multilayer perceptron, DL methods have greatly enhanced feature extraction, task accuracy, and scalability with larger datasets. Unlike the other machine learning methods, DL is based on artificial neural networks, with 'deep' referring to the depth of neural layers. The deep networks gradually enlarged their depth (hence complexity); correspondingly, they can fit bigger datasets. When training data has become more available in recent years, thanks to the development in the data acquisition sector, the performance of AI can be expected to improve.

A typical workflow for DL process (illustrated in Figure 1) includes the inference phase and a relatively long training process. Generally, a DL model must be first trained by feeding it a dataset; afterwards, it can be expected to make inferences from related unseen data.

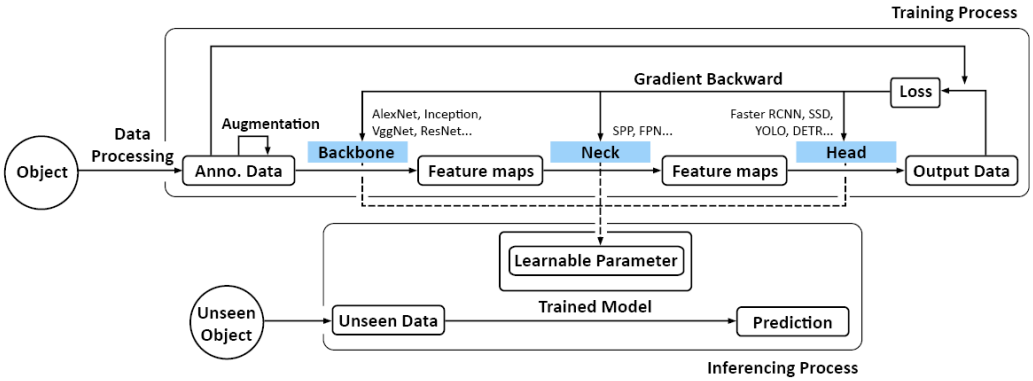

Figure 1. A typical pipeline of DL for object detection.

The training process of a DL model starts after the data acquisition and a series of processing, including cleaning, pre-categorising (structuring), annotation data augmentation, etc. The input data will go through the "backbone" network, a feature extractor. The extracted feature will pass through the "neck," which collects, combines, and transforms features. Then, the "head" will process the output and make decisions. The loss will be generated to start the optimization process, like gradient descent. Based on the loss function, the model will slightly update the parameters in the neural layers to reduce the loss. This process will run repetitively for numerous epochs until an acceptable loss value is achieved. This process could be time-consuming, but afterwards, the learnable parameter will be ready for prediction on unseen data, the so-called inferencing process. The input will go through the whole network once and not involve the gradient, no longer updating the learnable parameters.

1.2. Digitalization in Architectural Scenario: Data Collection, Fusion, and Processes

This paper intends to focus on the architectural scenarios, particularly CH, underlying how AI has been used in this field, supporting the documentation processes of existing buildings. Generally, different scenarios, for example, remote sensing or medical images, lead to different representation scales, error tolerance, and digitalisation methods. The characteristics of architectural spaces bring specific challenges to the practical applications:

 doi:10.20944/preprints202406.1162.v1

(1) Architectural elements are large and complex: architectural elements like pillars and walls cannot be represented by merely one shot of photo or scan in close range. The dataset that describes the elements of a big volume will require relatively more computer resources for storage and processing.
(2) Unavoidable noises and shadows: Architectural space is not an ideal space of simple geometric shape but a space of volumetric changes. Despite that, it is usually occupied with items casting shadows in 3D. Therefore, the digitalisation work of full coverage is costly and time-consuming.
(3) Objects of interest: the objects related to architectural elements, components, structures, and joints are featured with multiple scales of geometric features. Some objects, like walls, require global information. Some, like cracks and disintegration, need the millimetre-scale representation of the geometry. Highly decorated architectural elements, especially in the case of cultural heritage, are one of the most challenging objects because they are easily confused with some less important components or facilities.
(4) Recurrent monitoring in architectural spaces: The architectural scenario is close to human activity: from daily residential usage to massive production, architectural spaces vary in functional typology over time. Nonetheless, architecture is constantly threatened by environmental issues, anthropogenic damages, and continuous interventions. So, data acquisition is a constant need for monitoring. The acquired data will be used for object tracking, detecting translation, deformation, twisting, chromatic alteration, etc.

Understanding the state is crucial for ensuring construction and lifelong protection, and a reality-based digitalisation process is the first mandatory step in this understanding process. Documentation is always a key topic in architectural and building scenarios, starting from the needs derived from the construction and conservation practices for historical heritage.

The abundant bibliography of the last twenty years shows that research related to documentation in architectural and building scenarios covers all the phases of digitalisation, from field survey methods and data elaboration procedures to representation and fruition methods. The high demands of as-is documentation and the complicated semantic structures for architectonic elements and facilities stimulated the research and practical applications.

Regular maintenance, restoration, and investigation missions yield valuable data over time, especially in Cultural Heritage (CH) scenarios. However, the lengthy intervals between documentation updates often result in outdated measurement and representation technologies: traditional manual methods and paper-based drawings of 30 years ago are being replaced today by digital reality-capturing techniques, enhancing the modern documentation process. Various ways have been used to record architecture, such as writings and drawings in the past centuries. From the 19th century, photography and phonautograph emerged, producing photos, audio, and video records. Yet 2D photography was not enough for architectural projects; they are better supported with 3D information [7]. Photogrammetry and laser scanning techniques have provided a boost to 3D nowadays and have become the most favoured techniques for reality-capturing activities in architectural scenarios. In particular, image-based 3D reconstruction techniques (formerly called photogrammetry) have become very popular over laser scanning methods in the last few years because of their lower cost and flexibility. Also, photogrammetry, as a competitor to modern range-based mobile mapping systems, even if not yet largely tested in the architectonic and construction field, is used in mobile modality and gives quite good results in terms of reliability. Achille, Fassi and Fregonese, 2012 addressed their practices in the Milan Cathedral [8], Perfetti, Polari and Fassi, 2017 have even presented fisheye photogrammetry in narrow spaces [9]. The photogrammetric technique can generate spatial information from paired images and be applied to IRT [10] and multispectral [11] images.

Though not yet widely seen applications in the architectural digitalization field, but quite relevant in AI applications, RGB-D must be mentioned for its capability to integrate real-time 3D data and images[12]. This 2D representation of the 3D has become an object of research in computer graphics and computer vision [13,14]. Additionally, in the 3D geometric acquisition, other types of investigations can be integrated with geometrical ones to better understand the heritage and its characteristics. As presented by Adamopoulos and Rinaudo, 2021 [15] infrared thermography (IRT),

 doi:10.20944/preprints202406.1162.v1

multispectral imaging, ground penetrating radar (GPR), and active elastic wave techniques (Sonic and ultrasonic sensing techniques) are investigations that allow us to surpass the limit of human eyes, helping the identification of pathological issues, material differences, the presence of damage, or the changes in the material's physical properties behind the surface [16].

In addition to digitization, the data utilization methods, i.e., how the acquired data are made available to operators to be read, processed, and interpreted, are equally important. This is probably the most critical aspect nowadays, especially in architecture and cultural heritage. As a matter of fact, there are no standard methods or technologies for data storage and data sharing nor efficient and reliable automatic methodologies to aid data interpretation and processing.

It is precisely here that AI techniques can help in the future. Dealing with architectural topics means dealing with a large amount of data, and the manual approach has become time-consuming and unacceptable. Sharp techniques are urged to help human intervention and facilitate greater objectivity in analysis and results. Recent years have seen many attempts to integrate AI in architectural digitalization scenarios. As can be expected, AI tools are not initially developed for this particular purpose, and as is normal, the "needs emerge in the practical scenes." AI algorithms used in other fields are mainly tested by applying specific adjustments to satisfy the requirements, like inference speed, wide-range categories, noise robustness, and training efficiency.

For example, AI tasks that conduct automatic semantic classification and segmentation can make more accessibilities for investigators and monitors. As for images, a wide variety of DL models have been developed that could perform classifications such as VGG16 [17], Inception [18], and Residual Networks [19]. More complicated architectures were built to deal with object detection tasks, such as Yolo [6], FAST R-CNN[7], and semantic segmentation as FCN [20]. Multiple methods are developed to process point clouds, like random forest [9–12], support vector machine (SVM) [13], PointNet [14,15], and other machine learning methods in post-processing.

Many researchers have been testing the AI processing of 2D and 3D data, and recently, methods that combine multiple techniques have also emerged. These methods were tested in a practical scene and adjusted to practical needs and available data acquisition solutions. Consequently, the updated techniques are expected to accelerate the construction and preservation activities. They could provide reliable assistance in assessing the general situation of the scene and finding objects with locations, contributing to 3D data interpretation and big-data management.

1.3. Data Types and AI Methods

1.3.1. 2D Data

2D photographs are the predominant kind of data utilized in the documentation process and are currently undergoing digitization. The photo has been prevalent in the sphere of "architectural reading", documentation, and representation, gradually replacing, to some extent, the traditional practice of sketching for documentation purposes. Since its inception in 1851, photogrammetric surveying has used photographs and techniques like triangulation to obtain 3D measurements. Over time, it has evolved into a crucial instrument for driving progress and development in this field. Currently, photographs serve as the foundation for photogrammetric surveys and are commonly utilized to incorporate additional data, such as point scanner data. The major types of photos often include high-resolution digital frame images captured by traditional photography devices and fisheye images obtained from inexpensive portable sensors, such as those placed on drones. Additionally, panorama images are commonly used to add colour to static and mobile scanner data.

2D data processing has been a heated topic in the computer vision field for a long time. The first known neural network application for classification tasks dates to 1989 [6]. The researchers made a large database (MINST) of the handwritten samples of 10 digits and built a network with limited layers to recognize which number is written. Afterward, the dataset comes into the research field. Dataset preparation became the basis of other research [21–23]. Depending on DL tasks, the enlarging size of the sample and the increasing number of categories stimulated related research, including topics like the scale of the dataset (referring to the number of categories and instances), the semantic

 doi:10.20944/preprints202406.1162.v1

hierarchy of the classes, accuracy (reliability of the annotation), and diversity (appearance, positions, viewpoints and so on) etc. Works have also addressed the issue of "things" (objects with a well-defined shape, e.g., cat, person) and "stuff" (amorphous regions, e.g., sky, forest). The research emphasised the importance of stuff and discussed the contextual correlation to things [24].

Many image classification models emerged in the past years. After LeNet, typical man-crafted networks with limited layer depth were developed, such as VGG [17], inception network [18], etc. Residual network, known as ResNet [19], came out in 2015. It introduced the concept of residual connection, solved the problem of gradient vanishing, and allowed layer depth increases in the latter model.

At the same time, the object detection task was attracting attention from the researchers (**Error! Reference source not found.**). Classification merely defines the image; object detection attaches semantic meanings to the pixels or point locations, hence more useful information. An initial application for face detection [25] applied binary classification on sliding windows. The latter trend integrates hand-engineered feature extractors [26–28] with machine learning methods. The solution for object detection is later developed into two mainstream approaches: the two-step approach, which first finds where the objects could be and then classifies, and the one-step approach, which integrates localisation and classification in one pass. In recent years, approaches that got inspiration from NLP field also came into sight.

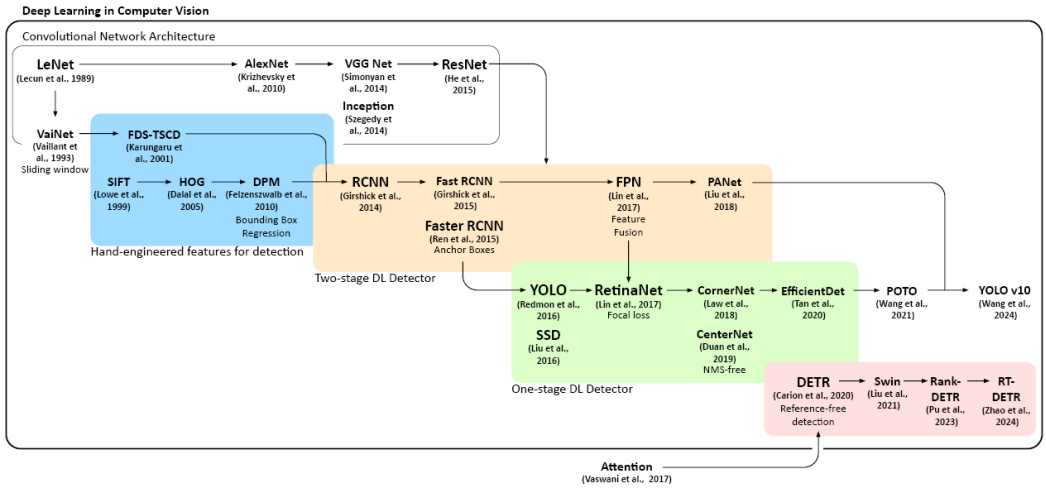

Figure 2. Chronological overview of 2D image object detection algorithms

In 2013, R-CNN was introduced by Ross Girshick et al. They innovatively applied a selective search algorithm, which extracts regions of interest (RoI) for later classification. The selective search turned out to be time-consuming. To make it fast, in Fast R-CNN [29], the input image is first processed by a neural network. Afterwards, the feature map will be cropped by region and passed through the prediction head. The Faster R-CNN [30] made the process faster by introducing a separate network to predict the region proposals. It uses the concept of anchor that predefines the bounding boxes and reshapes them using RoI-based (Region of Interest) methods, then the output goes to the prediction head. Algorithms like Faster R-CNN are considered a typical two-step approach; they will first make region proposals and then classify only proposed crops of images using convolutional networks.

Unlike the two-stage object detector, one-stage detectors use a fully convolutional approach in which the network can find all objects within an image in one pass. A famous example is YOLO (You Only Look Once) [31]. It is refreshingly simple, solely one convolutional network that outputs bounding boxes and classification probabilities at the same time. Different from that, SSD (Single-Shot Detector) [32] uses more feature layers to predict the boxes and the category confidences, allowing predictions at multiple scales (**Error! Reference source not found.**). CornerNet [33] introduced an innovative approach that, instead of defining the bounding box as the x and y coordinates of the box centre with h and w, uses a pair of diagonal key points. CenterNet [34],

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 June 2024 doi:10.20944/preprints202406.1162.v1

presented by Duan et al., 2019 detects objects using centre key points and corners [33]. They addressed a common defect of all one-stage approaches, as networks cannot pay attention to internal information within the cropped region without RoI extraction.

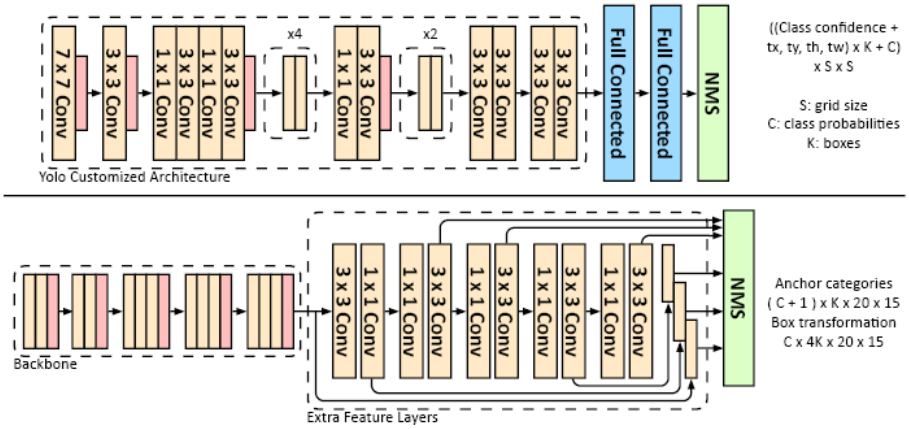

Figure 3. A comparison between two single shot detection models: SSD and YOLO.

DETR (DEtection TRansformer) [4] uses a "transformer network" to perform both feature extraction and object detection simultaneously. The transformer network was initially used for sequence modelling in natural language processing tasks, but it has also recently been applied in computer vision. DETR uses the "attention mechanism" to adaptively focus on the important parts of an object without requiring prior bounding boxes or anchor boxes. The performance of the COCO dataset can compete with that of an optimized Faster R-CNN. Later transformer-based approaches can be seen in [34–36]. They further mitigate the gap between NLP and CV.

Currently, the available architectures of object detection are different in terms of approaches (one-step, two-step, transformer), convolutional feature extractors (VGG, Inception, ResNet), and detection heads. The requirements for training time and computational resources vary, as well as the performances. Huang et al., 2017 [37] have discussed the speed and accuracy issues. Later detection models surpass the performances of the previous, but some facts remain: without region proposals, one-step detectors require lower computational resources and, hence, relatively faster. Later approaches (e.g. CenterNet) further accelerate the process by avoiding Non-Maximum Suppression (NMS). However, in terms of overall mean Average Precision (mAP), one-step can hardly catch up with two-step approaches.

When it comes to practical scenarios, in architectural scenarios, the detector's requirement is mainly the computational resource during the in-situ data acquisition process, as the ex-situ post-processing (e.g., 3D reconstruction) is supported with time and computational resources. In this case, one-stage detectors like YOLO are the preferred choices. As for the CornerNet and later CenterNet, their performances of the irregular shape are critical [34]. DETR is a late promising solution, but the complicated training process [4] makes it unsuitable for case-wise application in architectural fields.

Object detection architectures are the basis of many downstream applications, such as semantic segmentation, human key point detection, and image captioning. In 2017, Kaiming He et al. introduced Mask R-CNN [38] as an extension of Faster R-CNN. They added a branch to the network that predicts each object's segmentation mask and its class and bounding box. This allowed the model to perform object detection and instance segmentation with state-of-the-art accuracy.

1.3.2. 3D Data

Currently, 3D data serves as the foundation for digitalization as it is crucial to accurately and comprehensively describe the geometry of objects, particularly in the fields of architecture and cultural heritage.

The primary digital data result is 3D point cloud data, comprising a collection of 3D points depicting an object's outer surface geometry or environment. A point cloud obtained from a scanner is typically generated in real time, but a point cloud obtained from photogrammetry may require

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 June 2024 doi:10.20944/preprints202406.1162.v1

more time to be created. Many processes, including registration, referencing, filtering, down-sampling, and feature extraction, are often carried out during post-processing and are only partially automated. Specifically, extracting features required for a subsequent modelling or architectural restitution phase and data interpretation still rely on manual operations performed by individual operators. Moreover, one of the main challenges of point cloud data is its high dimensionality, which makes it difficult and time-consuming to manually process and analyse a large amount of data representing complex geometry.

For these reasons, segmentation and classification are the two most discussed and required AI applications because, based on those results, point clouds can be successfully exploited and better comprehended [39].

AI, especially the DL approach, has also shown great interest in processing 3D point cloud data (**Error! Reference source not found.**). In a supervised machine learning approach, semantic categories are learned from a manually annotated sub-set of data, which are used to train the classification models like SVM, RF, and NB. The input point cloud must pass through hand-crafted feature extractors to compute individual points' close and global neighbouring features. After the model is trained, it will be used to conduct the semantic classification of all the points in the dataset. With proper features, machine learning models can be effective in specific tasks even with limited annotated data.

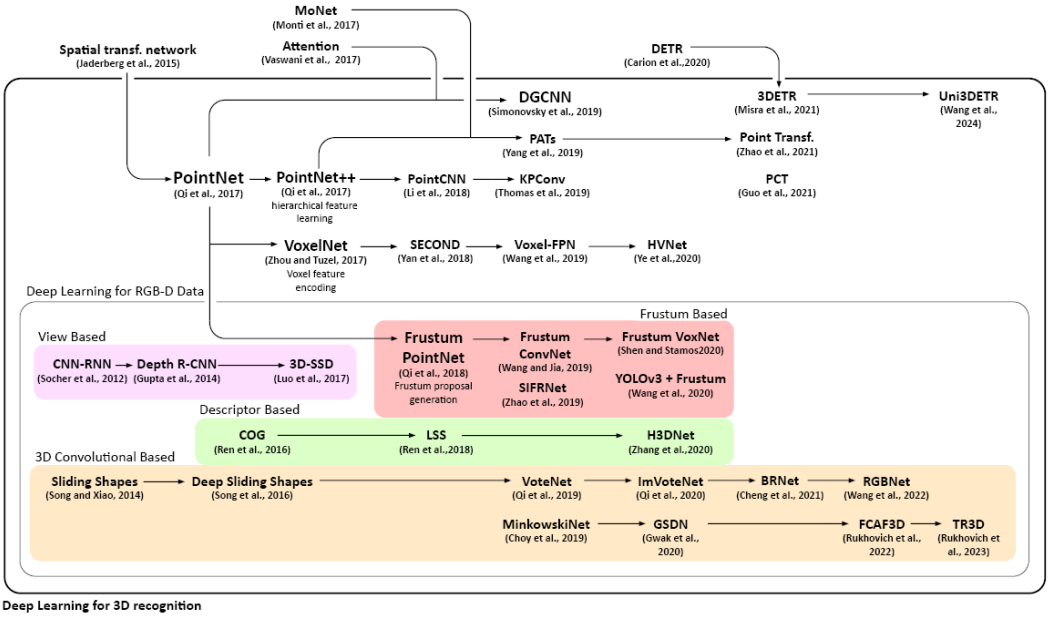

Figure 3. Chronological overview of 3D object detection algorithms

Among the different DL approaches, the feature learning methods are initially divided into point-based and tree-based approaches [40]. The first directly takes the raw point cloud as the input for training the DL network. The second employs a k-dimensional tree (Kd-tree) structure to transform the point cloud into a regular representation (linear representation afforded by the group action), and then feeds this into DL models. The concept of using representatives saves computational resources.

In 2017, PointNet [3] was introduced by Charles et al. It is a DL architecture that directly learns point-wise features from unordered point clouds without the need for pre-processing such as voxelization or projection onto 2D grids. It has shown impressive performance in point cloud classification tasks. Noticeably, they added a data-dependent spatial transformer network to canonicalize the data. However, the PointNet could not capture local structures, which limited its ability to recognise high-granularity patterns and its generalization ability to large and complex scenes. PointNet++[15] upgraded the previous network to address this defect: a hierarchical structure was designed. The new architecture is composed of several set abstraction levels. These abstraction

 doi:10.20944/preprints202406.1162.v1

layers aggregate multi-scale information according to local point densities, making the learning process efficient and robust.

PointCNN [41] by Li et al., 2018, proposed an X-transformation to address the issue of point-wise feature weighting and permutation of the unordered point clouds. Another point-wise convolution approach is inspired by image-based convolution. KPConv [42] by Thomas et al., 2019, uses Kernel Point Convolutions to process point clouds without intermediate representation. It has shown impressive performance in point cloud segmentation tasks.

The attention transformer approach is used for 3D recognition. PATs (Point Attention Transformer) [43] represent points by position and neighbourhood and learn features through Multi-Layer Perception (MLP). Later developments continued in PCT (point cloud transformer) [44], Point transformer [45], 3DETR [46], and Uni3DETR [47] have also achieved state-of-the-art results.

DGCNN [2] by Wang et al., 2019, or Dynamic Graph CNN is also inspired by PointNet. It uses a graph neural network to model the local geometric structures of point clouds and dynamically constructs a graph based on the spatial relationships between points. It then applies a series of graph convolutional layers to learn hierarchical representations of the input point cloud, followed by a max pooling operation to obtain a global feature vector. Finally, a fully connected network is used to classify the point cloud.

Zhou and Tuzel, 2017, proposed VoxelNet [48], which eliminates manual feature engineering of the point cloud. It unifies a deep convolutional network as a feature extractor and region propose network into a single-stage, end-to-end DL network by using the Voxel Feature Encoding (VFE) concept. The encoder is widely applied to many later models, like SECOND [49], Voxel-FPN [50], and HVNet[51].

1.3.3. RGB-D Data

RGB-D data are 2D images where each pixel includes colour information (RGB) plus corresponding depth information. This data type can be achieved using depth sensors, like structured light and Time-of-Flight (ToF) cameras. RGB-D data can also be achieved from the photogrammetric matching process that allows the estimation of the depth maps from oriented RGB images in the space. These algorithms compare the disparities between images taken from different angles to calculate depth. An example of this is the Semi-Global Block Matching (StereoSGBM) algorithm. This OpenCV implementation uses semi-global matching (based on the work of StereoSGM [52]) for stereo images to produce accurate depth maps.

Thus, RGB-D data is used more for 3D interior reconstruction, especially semantic VSLAM direction [53]. However, the mentioned cameras are light-sensitive, can produce noisy, low-resolution models, and have problems with narrow horizons. This type of data is mentioned due to its growing usage and the fact that it works better for 3D object detection than other formats [14].

In deep learning, convolutional networks were introduced to compute matching fields between two images [54]. Afterwards, research on monocular depth estimation addressed the high costs and complexity associated with acquiring data from multiple views. Algorithms are made to infer detailed 3D structures from single still images [55] using Markov Random Field (MRF), and convolutional neural network [56]. End-to-end tools have been developed to synthesise novel views directly from images. One example is the Deep3D model [57], which uses a convolutional neural network to generate the corresponding right view from an input source image.

Methods used to classify and localise objects using RGB-D data can be categorized (**Error! Reference source not found.**) as follows:

(1) **View-based:** Some researchers make use of RGB-D data by considering the depth map as an additional channel [58]. Some process the 3D data as front-view images [59,60], or projects 3D information to bird's view [61]. Depth R-CNN[62] was inspired by the 2D object detection model RCNN, it introduced geocentric embedding for better usage of the depth maps. Similarly, 3D-SSD [63] is a 3D generalization of the SSD framework.

(2) **Frustum-based** (2D driven 3D): Frustum-PointNet [64] as extensions of PointNet turns to process RGB-D information. It extracts the 3D bounding frustum of an object by projecting 2D bounding boxes from image detectors to 3D spaces. Then, within the trimmed 3D, segmentation and box

 doi:10.20944/preprints202406.1162.v1

regression are consecutively performed using variants of PointNet. It has addressed some of the limitations, such as local features handling, rotation sensibility, feature extraction, using hierarchical neural network architecture and graph convolutional network. Frustum-PointNet was the breakthrough method at time, it inspired later models such as YoloV3 & F-PointNet [65] and Frustum Voxnet [66].

(3) **Descriptor-based:** Geometric descriptors for 3D object detection were introduced. COG [67] links the 2D appearance and the 3D pose of object categories. Ren et al. also introduced latent support surfaces [68], where the location can explain shape variation in 3D.

(4) **Convolution-based:** Sliding Shapes [5] applies a 3D sliding window to detect 3D places directly. It was later updated in Deep sliding shapes [69], introducing a 3D region proposal network to speed up the computation. Qi et al. presented VoteNet[70], which directly detects objects in the point cloud. It has addressed the challenge that the centroid of an object in 3D can be far from the surfaces by applying Hough voting.

2. AI Applications in Architectural Scenarios

2.1. 2D Applications

AI applications with 2D input typically involve training models for classification, object detection, and segmentation tasks. Table 1 lists recent 2D AI applications in the architectural scenarios. Researchers were testing them in specific scenes, like construction field, material defect detection. YOLO and Faster RCNN are the most used models due to their effectiveness and relatively simple training process. These models are often seen tested with modification to their architectures. In current stage, 2D applications achieve promising results with limited categories for detection. Considering the models can perform tasks properly, some researchers projected or interpolate 2D semantic information to 3D. Additionally, AI models can be used for tasks like photo matching, orientation, and image masking, which related to advanced photogrammetric approach.

Table 1. 2D AI applications tasks comparison

Authors	Scene	Data Type	Labels	Trainset	Methods	Evaluation	3D
Hatır et al., 2021	Stone heritage	Photo	9	Subset from site	Mask RCNN	mAP 0.98	-
Mishra et al., 2022	Heritage exterior	Photo	4	Subset from site	YOLO v5	mAP 0.93	-
					Faster RCNN	mAP 0.85	
Liu et al., 2022	Constructi on filed	Photo	3 separatel y	Subset from dataset	Faster RCNN	mAP 0.90, 0.87, 0.74	-
					SSD	mAP 0.88, 0.69, 0.53	
Kwon et al., 2019	Stone	Photo	4 separatel y	Subset from dataset	Faster RCNN *bb. Inception	Avg. conf. 0.95	-
Wang et al., 2019	Mansory facade	Orthopho to	2	Subset from dataset	Faster RCNN *bb. ResNet101	mAP 0.95	-
Idjaton et al., 2022	Stone facade	Photo	1	Subset from site	YOLO v5 & Attention	mAP 0.79	-
					YOLO v5x	mAP 0.74	
Karimi et al., 2023	Bridge defect	Photo	7	Individual case	Inception-ResNet-v2	Accu. 0.96	-

 doi:10.20944/preprints202406.1162.v1

Guerrieri et al., 2022	Stone pavement	Photo	4	Subset from site	YOLO v3	Accu. 0.91-0.97	-
Zhang et al., 2024	Pathology	Photo	12	Subset from dataset	ResNet	Accu. 0.72	-
					YOLO v5	mAP_0.5 0.34	+
Pathak et al., 2021	Mansory heritage	Photo & Rendered	2	Individual case	Faster RCNN Nas	mAP 0.39	+
					Faster RCNN Resnet-FPN	mAP 0.58	
Grilli et al., 2018	Heritage exterior	Orthopho to	7	Subset from site	Random Forest etc.	Accu. 0.44-0.69	+
Grilli et al., 2019	Heritage interior	Panoram a	1	Subset from site	K-means clustering	Accu. 0.91	+

A Mask R-CNN application on stone heritage was tested [71], trained and evaluated on photos collected from the same site. Similar work was done by Mishra et al., 2022[72], who compared YOLOv5 with FasterRCNN(ResNet101) to detect architectural defects, like exposed bricks, cracks, and spalling. The result shows that YOLO surpasses the latter regarding maximum mean average precision. Liu et al., 2021 [73] compared Faster RCNN with Single Shot multi-box Detector (SSD) in the modular construction scene. The under-detecting objects are common modular objects, including modular panels, barricades, and site fences. Here, the Faster RCNN always outperforms the SSD in terms of average recall and mean average precision.

Following the instructions of the illustrated glossary on stone deterioration patterns [74], there have been many trials of using DL methods to detect stone face deterioration. Faster R-CNN is introduced to detect stone and masonry structure damages [75,76]. Idjaton et al. [77] used an enhanced YOLO network with a transformer encoder to process sub-images from orthophotos. The inception-ResNet model was used for defect detection on images taken from bridges[78].

The dataset used for the training can influence the effectiveness of machine learning models: either the training set is not representative, or lacks diversity, or is not properly processed and annotated, the model will not generalize well to unseen data. Guerrieri and Parla [79] used YOLOv3 for road pavement damage detection, addressing the issue with no available datasets for the research matter. They collected the data using a calibrated front camera device installed in a survey car. An interesting example is from Tatzel et al., 2018 [80], who introduced CNN into the laser-cut processing of stainless steel; the visualization of the CNN, using layer-wise relevance propagation (LRP)[81], offered a better understanding of the cutting edge. This approach enables the model to do the class-activation mapping. Unlike semantic segmentation, it provides heat maps that indicate the probable area on images that indicate the labels. A similar work evaluated DL model behaviour on architectural pathology [74]; their result suggested that object detection tasks can be challenging in architectural scenarios. Because, from the provided training set, the objects of interest cannot be easily defined with clear boundaries, the visual representations vary from the light environment and view perspective.

Nowadays, it is difficult for a 2D investigation to be sufficient if unrelated to a 3D localisation of information. For this reason, many examples in the literature show that the classification and detection results from 2D images can be interpolated back to 3D space. A practice introduced by Pathak et al., 2021[82] tested Faster RCNN architecture to detect cracks and spalling on unseen rendered images from a 3D model of Hampi (India) by using a labelled database from Wat Phra Si Sanphet Temple (Thailand). By this method, the classification result can be mapped onto the model, getting the location of the detected damages. Another approach, "texture-based" classification [83,84] classifies 2D data unwrapped from 3D models, projecting them onto 3D geometries for better understanding. The classification is performed on the texture image and orthoimages obtained from

 doi:10.20944/preprints202406.1162.v1

the model. The results are then reprojected on the 3D model. Optimized models, orthoimages, and UV maps are created for each case under study.

The output of DL models that process 2D images can benefit further 3D reconstruction. In the same logic, based on global and accurate geometric information, the classification results from the point cloud can be interpolated back to images for further predictions. A potential research field seeks to integrate 2D DL processing with photogrammetry, the technology that can obtain reliable spatial (3D) information by processing photographic images (2D).

An important image process in photogrammetry is image matching, which generates pixel positions in multiple views for camera relative pose estimation. The advent of feature extractors, such as SIFT (Scale-Invariant Feature Transform) [85], SURF (Speeded Up Robust Features) [86], ORB (Oriented FAST and Rotated BRIEF) [87], largely accelerated this traditionally manual process of image matching. Subsequently, some research has delved into using neural networks for keypoint detection and description [88–90], further advancing the field.

Stathopoulou and Remondino (2019) have discussed this semantic photogrammetry application [91], stating that the 3D reconstruction can be optimized if semantic labelling is provided in 2D data. In this work, they applied a convolutional neural network to process 2D images of the historical façade dataset, and the label was reused in 3D reconstruction and later processing. The semantic segments are tested to have constrained the tie points extraction, reinforced image matching, reduced noise and, most importantly, indirectly labelled the generated 3D data. In the later work [92], they applied semantic constraints during the depth map fusion step in 3D reconstruction (see **Error! Reference source not found.**). The automation is boosted by selectively obtained segmented point clouds for each label. This study proved that this method can improve the Multi-View Stereo (MVS) reconstruction [93] in terms of filling the information gaps and preserving details. Work from another team [94] applied a DeeplabV3+ model to detect architectural and other elements from photos and remapped the refined semantic segmentation to a pre-constructed 3D model.

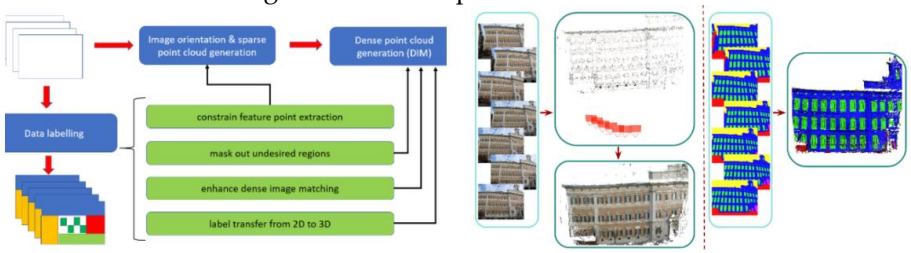

Figure 6. The proposed semantic photogrammetric pipeline in the work of Stathopoulou et al. 2019.

Another promising approach is integrating DL processing with Simultaneous Localization and Mapping (SLAM) technology. The SLAM problem involves constructing a map of the sensor environment and keeping track of its location. Visual SLAM utilizes photos to reconstruct the surroundings in real-time. It can be integrated with AI techniques to accelerate its process. The integration detects objects of interest and outputs corresponding 3D localizations.

Semantic SLAM was brought up in 2011 by Civera et al. [95]; their work combined object recognition that inserted precomputed known objects into the monocular SLAM map. In 2018, a semantic SLAM system in Dynamic environments (DS SLAM) [96] based on ORB-SLAM2 was an example that used RGB-D data as input. It combines a real-time semantic segmentation network, SegNet, to filter out dynamic portions of the scene. This integration improved the robustness and accuracy of the SLAM performance in moving scenes. Truong et al., 2020 [97] proposed a new semantic visual simultaneous localization and mapping system that integrates a visual SLAM system using an RGBD camera. With the YOLO neural network application, this system can detect objects when the device moves in the environment, incorporating semantic information for the building process and visual navigation using object-based landmarks.

2.2. 3D Application

 doi:10.20944/preprints202406.1162.v1

Machine learning methods are generally favoured in 3D applications, especially in CH, where the wide variety of geometries and forms that cannot be easily collected to a pre-set that is complete and usable for DL training. For a case (usually small) that requires automatic segmentation and classification, though annotation is based on specific needs, the effort is fast and efficient for training machine learning models. Researchers tried to apply the DL method, in which the model can be trained on a bigger dataset that covers more cases. However, trained models for specific tasks cannot easily generalize to unseen datasets. Therefore, the DL model can be practical in detecting objects of more general interest.

Table 2 lists recent 3D AI application in different scenes, including urban, heritage and architectural contexts. Using subset of the complete dataset for training is a favoured and effective approach because it guarantees the consistency of the data characteristics. There is still 2 trends for features: learnable feature extractors and human engineered ones. The listed cases also underline the common burden of big data size in practical 3D scenes.

Table 2. 3D AI applications tasks comparison

Authors	Scene	Trainset	Methods	Feature	Dataset	Labels	Precision
Sun et al., 2018	Urban	Subset from site	RF on supervoxel	Covariance features	-	8	Ov. Acc. 0.92
Grilli et al., 2019	Heritage exterior	Subset from site	RF	Covariance features	2.2, 1.1 million	10, 14	Avg. F1 0.92, 0.82
Teruggi et al., 2020	Heritage complex	Subset from site	RF	Covariance features	3.6 billion	25	F1 0.94-0.99
Zhang et al., 2022	Heritage complex	Subset from related site	RF	Covariance features	354.2 milion	18	F1 0.92-0.97
					24.2 million	18	F1 0.92-0.99
					156 million	19	F1 0.90-0.99
Grilli et al.,2020	Heritage exterior	Individual case	RF	Covariance features	6-14 million	9	Acc. 0.78-0.98
Cao et al., 2022	ArCH	Subset from dataset	RF & DGCNN	Covariance features and none	14, 10, 17million	9	Ov. Acc. 0.62-0.98
Malinverni et al., 2019	ArCH	Subset from dataset	PointNet++	-	25-569 million	4	Wgt. Avg. F1 0.31
Pierdica et al., 2020	ArCH	Subset from dataset	Modified DGCNN	HSV + Norm	Avg. 114 million	10	Mean Acc. 0.74
Matrone et al., 2020	ArCH	Subset from dataset	Modified DGCNN	Covariance features	Avg. 114 million	10	Wgt. F1 0.82-0.91
Cao et al., 2022	ArCH	S3DIS (interior)	PointNet	-	Avg. 114 million	10	mIoU 0.31
			KPConv				mIoU 0.65
			DGCNN				mIoU 0.47
Cao et al., 2020	ArCH	Subset from dataset	3DLEB-Net	-	Avg. 114 million	10	Ov. Acc. 0.67-0.77

 doi:10.20944/preprints202406.1162.v1

Ma et al., 2020	S3DIS	Synthetic data from dataset	DGCNN	-	695+ million	12	Avg. Prec. 0.53-0.65
Landrieu and Simonovsky, 2018	S3DIS	Subset from dataset	Superpoint Gragh + GCNN	-	695 million	12	Ov. Acc. 0.86
Li et al., 2019	S3DIS	Subset from dataset	ResGCN-28	-	695 million	12	Ov. Acc. 0.86
Hu et al., 2020	S3DIS	Subset from dataset	RandLA-Net	-	695 million	12	Ov. Acc. 0.88

An example of a machine learning approach that works on the point cloud model is presented in [98]. This approach uses the RF algorithm and works on a set of manually labelled samples with computed geometric covariance features. The model to be trained is fed with a manually defined training set, which takes up 50% of the whole dataset. Then, it generates predictions on the previously segmented evaluation set (the remaining 50%) to calculate its performance. This work introduced a supervoxel-based local context to reduce the computation cost since the voxelization process is considered a sub-sampling process.

The RF model can directly work on each point of the point cloud. In an early practice [99], the dataset was subsampled to accelerate the process. Provided with covariance features in the training process, the trained RF model has achieved promising results. Inspired by this work, an MLMR approach is introduced [100]. It works hierarchically on specific portions of the whole dataset, classifying it at different resolutions with increasing detail as the level of classes increases; this method has proved to be computationally unexpensive and has allowed higher accuracy to be achieved on more complex architectural heritage buildings. Initially, the dataset is subsampled to a lower resolution (depending on the dimensions of the considered case study). Training a specific RF classifier, big macro-elements are classified. The result is then back interpolated on a point cloud of higher resolution to subdivide the elements that require higher geometric accuracy. The process iterates up to the classification of the full-resolution dataset (initial resolution) (**Error! Reference source not found.**). In contrast to the non-hierarchical approach, specific RF models are trained for each classification, but only a small number of labelled samples are required. The data are hierarchically split into sub-classes while the level of geometric detail increases, allowing the discernment of architecture components processed on a limited portion of the dataset at a relevant resolution. The validity of the approach has been previously proved in Chinese wooden architectures [101]. The application of a single machine learning model across large and variable architectural datasets has been extensively tested [102], but the generalization ability in the hierarchical approach requires further discussion [103,104].

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 June 2024 doi:10.20944/preprints202406.1162.v1

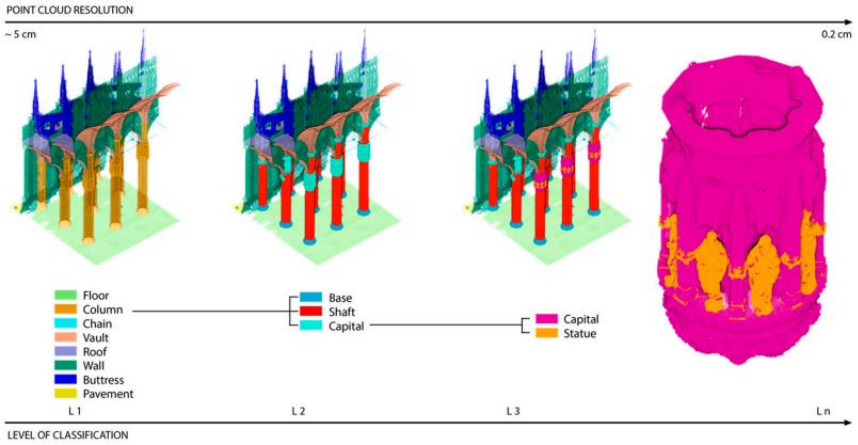

Figure 5. MLMR classification levels (till capital details) for the Milan Cathedral by Teruggi et al., 2020.

In contrast to machine learning approaches, and based on the prospering of point cloud datasets, recent years have seen the application of DL networks to point model classification; examples include PointNet, PointNet++ [105], PCNN [106], and DGCNN [2]. DL frameworks based on these neural networks have been applied to the digital Cultural Heritage domain. Consequent applications, such as the improved DGCNN [107,108], support features such as normal and HSV colours coupled with the points' x, y, and z coordinates. The ArCH (Architectural Cultural Heritage) dataset [109] has been used to gather different pre-classified examples to train the network. A limitation is that the number of classes must be predetermined and constant across all cases in the training dataset. However, due to the uniqueness and vastness of different heritage building examples, it is still very difficult, if not impossible, to define a dataset with an adequate number of pre-classified examples to cover a complete range of heritage buildings. Some studies have used transfer learning [110] against this deficiency. Using a pre-trained example, they developed a DL approach that can reduce the need for a bigger pre-segmented dataset, obtaining encouraging results [111–113].

Another noticeable dataset is the "2D-3D-Semantics" (2D-3D-S) dataset [114], which provides mutually registered modalities from 2D, 2.5D, and 3D domains, including RGB, semantics, surface normal, and depth. Stanford's large-scale 3D Indoor Spaces Dataset (S3DIS) is included in this dataset, providing 695 million coloured points of interior scenes. [115] has used the S3DIS to generate synthetic data, and the augmented data was tested on DGCNN. Another research used a "SuperPoint Graph" (SPG) representation to encode the contextual relation between object parts. It uses a graph convolutional network and has achieved promising results without too much computational resources [116]. ResGCN-28 [117] and RandLA-Net [118] have been tested on this dataset and have shown satisfying performances.

3. Conclusion

This paper discusses numerous state-of-the-art works that use AI methods to support reality capture applications, mainly in architecture and cultural heritage. In general, reality capturing is an activity that uses multiple technologies to digitalize 3D environments. The different acquisition approaches and data formats largely shaped the later elaboration process, data interpretation and use, and methods for integrating AI. 2D and 3D data have advantages over others in architectural scenarios and often need to be supplemented to achieve a more comprehensive, multiscale degree of knowledge. Point clouds are suitable for representing accurate geometric information of large-scale elements, like architectural elements and large 3D environments. Things with small geometric features like cracks and nails are more recognizable with high-resolution images.

Thanks to the advanced research progress of deep learning networks that deal with image analysis, quick and autonomous assessment can be implemented in the data acquisition phase, which benefits photogrammetry, RGB-D-based approaches, and 3D point cloud processing. The literature

 doi:10.20944/preprints202406.1162.v1

shows that DL methods are more usable for 2D data, while machine learning is preferred for each 3D case. This could be explained by the shortage of point cloud datasets, considering that their collection, data processing, and annotation require much more effort than photos in terms of time, cost, and availability.

Vigorous research on 3D tasks allows for increasingly faster performance on 3D point cloud classifications and accurate localization. Though common elements of interest with regular shapes can be recognized well, DL integration in the reality-capturing field cannot yet be used and generalized for real practical scenarios.

Research has shown that integrating AI's understanding of both 2D and 3D data can significantly accelerate 3D reconstruction processes, especially in photogrammetry and the 3D semantic classification of point cloud models. Future research aims to enhance this integration for detailed, multiscale documentation of buildings and their conservation state. Practical tests are needed to evaluate the effectiveness of RGB-D-based approaches in architectural scenarios. Standardized methods for data fusion and related datasets are demanded. Further works is needed to build the bridge from complicated investigation and documentation activities to the Deep Learning techniques that facilitate human intervention in 2D images and 3D point cloud.

Author Contributions: Conceptualization, K.Z. and F.F.; methodology, K.Z. and F.F.; software, K.Z.; validation, K.Z. and F.F.; formal analysis, K.Z. and F.F.; investigation, K.Z. and F.F.; resources, F.F.; data curation, K.Z.; writing—original draft preparation, K.Z.; writing—review and editing, K.Z. and F.F.; visualization, K.Z.; supervision, F.F.; project administration, F.F.; funding acquisition, K.Z. and F.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the China Scholarships Council, grant number 202208520007.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author.

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

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-Time End-to-End Object Detection 2024. 10.48550/arXiv.2405.14458.
2. Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S.E.; Bronstein, M.M.; Solomon, J.M. Dynamic Graph CNN for Learning on Point Clouds. *ACM Trans. Graph.* **2019**, *38*, 146:1-146:12. 10.1145/3326362.
3. Charles, R.Q.; Su, H.; Kaichun, M.; Guibas, L.J. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); IEEE: Honolulu, HI, 2017; pp. 77–85. 10.1109/CVPR.2017.16.
4. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers 2020. 10.48550/arXiv.2005.12872.
5. Song, S.; Xiao, J. Sliding Shapes for 3D Object Detection in Depth Images. In Proceedings of the Computer Vision – ECCV 2014; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer International Publishing: Cham, 2014; pp. 634–651. 10.1007/978-3-319-10599-4_41.
6. LeCun, Y.; Boser, B.; Denker, J.; Henderson, D.; Howard, R.; Hubbard, W.; Jackel, L. Handwritten Digit Recognition with a Back-Propagation Network. In Proceedings of the NIPS; 1989.
7. Fassi, F.; Campanella, C. From Daguerreotypes to Digital Automatic Photogrammetry. Applications and Limits for The Built Heritage Project. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2017**, *XLII-5-W1*, 313–319. 10.5194/isprs-archives-XLII-5-W1-313-2017.
8. Achille, C.; Fassi, F.; Fregonese, L. 4 Years history: From 2D to BIM for CH: The Main Spire on Milan Cathedral. In Proceedings of the 2012 18th International Conference on Virtual Systems and Multimedia; 2012; pp. 377–382. 10.1109/VSMM.2012.6365948.
9. Perfetti, L.; Polari, C.; Fassi, F. Fisheye Photogrammetry: Tests and Methodologies for The Survey of Narrow Spaces. In Proceedings of the The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences; Copernicus GmbH, 2017; Vol. XLII-2-W3, pp. 573–580. 10.5194/isprs-archives-XLII-2-W3-573-2017.

 doi:10.20944/preprints202406.1162.v1

10. Patrucco, G.; Gómez, A.; Adineh, A.; Rahrig, M.; Lerma, J.L. 3D Data Fusion for Historical Analyses of Heritage Buildings Using Thermal Images: The Palacio de Colomina as a Case Study. *Remote Sens.* **2022**, *14*, 5699. 10.3390/rs14225699.
11. Raimundo, J.; Lopez-Cuervo Medina, S.; Aguirre de Mata, J.; Prieto, J.F. Multisensor Data Fusion by Means of Voxelization: Application to a Construction Element of Historic Heritage. *Remote Sens.* **2022**, *14*, 4172. 10.3390/rs14174172.
12. Lehmann, E.H.; Vontobel, P.; Deschler-Erb, E.; Soares, M. Non-Invasive Studies of Objects From Cultural Heritage. *Nucl. Instrum. Methods Phys. Res. Sect. Accel. Spectrometers Detect. Assoc. Equip.* **2005**, *542*, 68–75. 10.1016/j.nima.2005.01.013.
13. Zollhöfer, M.; Stotko, P.; Görlitz, A.; Theobalt, C.; Nießner, M.; Klein, R.; Kolb, A. State of the Art on 3D Reconstruction with RGB-D Cameras. *Comput. Graph. Forum* **2018**, *37*, 625–652. 10.1111/cgf.13386.
14. Wang, Y.; Wang, C.; Long, P.; Gu, Y.; Li, W. Recent Advances in 3D Object Detection Based on RGB-D: A Survey. *Displays* **2021**, *70*, 102077. 10.1016/j.displa.2021.102077.
15. Adamopoulos, E.; Rinaudo, F. Close-Range Sensing and Data Fusion for Built Heritage Inspection and Monitoring—A Review. *Remote Sens.* **2021**, *13*, 3936. 10.3390/rs13193936.
16. Orbán, Z.; Gutermann, M. Assessment of Masonry Arch Railway Bridges Using Non-Destructive In-Situ Testing Methods. *Eng. Struct.* **2009**, *31*, 2287–2298. 10.1016/j.engstruct.2009.04.008.
17. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition 2015. 10.48550/arXiv.1409.1556.
18. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions 2014. 10.48550/arXiv.1409.4842.
19. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition 2015. 10.48550/arXiv.1512.03385.
20. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation 2015. 10.48550/arXiv.1411.4038.
21. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition; 2009; pp. 248–255. 10.1109/CVPR.2009.5206848.
22. Everingham, M.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. 10.1007/s11263-009-0275-4.
23. Lin, T.-Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C.L.; Dollár, P. Microsoft COCO: Common Objects in Context 2015. 10.48550/arXiv.1405.0312.
24. Caesar, H.; Uijlings, J.; Ferrari, V. COCO-Stuff: Thing and Stuff Classes in Context 2018. 10.48550/arXiv.1612.03716.
25. Vaillant, R.; Monrocq, C.; Le Cun, Y. An Original Approach for The Localization of Objects In Images. In Proceedings of the 1993 Third International Conference on Artificial Neural Networks; 1993; pp. 26–30.
26. Dalal, N.; Triggs, B. Histograms of Oriented Gradients for Human Detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05); 2005; Vol. 1, pp. 886–893 vol. 1. 10.1109/CVPR.2005.177.
27. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D.; Ramanan, D. Object Detection with Discriminatively Trained Part-Based Models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1627–1645. 10.1109/TPAMI.2009.167.
28. Lowe, D.G. Object Recognition from Local Scale-Invariant Features. In Proceedings of the Proceedings of the Seventh IEEE International Conference on Computer Vision; 1999; Vol. 2, pp. 1150–1157 vol.2. 10.1109/ICCV.1999.790410.
29. Girshick, R. Fast R-CNN 2015. 10.48550/arXiv.1504.08083.
30. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks 2016. 10.48550/arXiv.1506.01497.
31. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection 2016. 10.48550/arXiv.1506.02640.
32. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In; 2016; Vol. 9905, pp. 21–37. 10.1007/978-3-319-46448-0_2.
33. Law, H.; Deng, J. CornerNet: Detecting Objects as Paired Keypoints 2019.
34. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows 2021. 10.48550/arXiv.2103.14030.
35. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs Beat YOLOs on Real-time Object Detection 2024. 10.48550/arXiv.2304.08069.
36. Pu, Y.; Liang, W.; Hao, Y.; Yuan, Y.; Yang, Y.; Zhang, C.; Hu, H.; Huang, G. Rank-DETR for High Quality Object Detection 2023. 10.48550/arXiv.2310.08854.

37. Huang, J.; Rathod, V.; Sun, C.; Zhu, M.; Korattikara, A.; Fathi, A.; Fischer, I.; Wojna, Z.; Song, Y.; Guadarrama, S.; et al. Speed/accuracy trade-offs for modern convolutional object detectors 2017. 10.48550/arXiv.1611.10012.
38. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN 2018. 10.48550/arXiv.1703.06870.
39. Grilli, E.; Menna, F.; Remondino, F. A Review of Point Clouds Segmentation and Classification Algorithms. *ISPRS - Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2017**, *XLII-2/W3*, 339–344. 10.5194/isprs-archives-XLII-2-W3-339-2017.
40. Liu, W.; Sun, J.; Li, W.; Hu, T.; Wang, P. Deep Learning on Point Clouds and Its Application: A Survey. *Sensors* **2019**, *19*, 4188. 10.3390/s19194188.
41. Li, Y.; Bu, R.; Sun, M.; Wu, W.; Di, X.; Chen, B. PointCNN: Convolution On X-Transformed Points 2018. 10.48550/arXiv.1801.07791.
42. Thomas, H.; Qi, C.R.; Deschaud, J.-E.; Marcotegui, B.; Goulette, F.; Guibas, L.J. KPConv: Flexible and Deformable Convolution for Point Clouds 2019. 10.48550/arXiv.1904.08889.
43. Yang, J.; Zhang, Q.; Ni, B.; Li, L.; Liu, J.; Zhou, M.; Tian, Q. Modeling Point Clouds with Self-Attention and Gumbel Subset Sampling 2019. 10.48550/arXiv.1904.03375.
44. Guo, M.-H.; Cai, J.-X.; Liu, Z.-N.; Mu, T.-J.; Martin, R.R.; Hu, S.-M. PCT: Point cloud transformer. *Comput. Vis. Media* **2021**, *7*, 187–199. 10.1007/s41095-021-0229-5.
45. Zhao, H.; Jiang, L.; Jia, J.; Torr, P.; Koltun, V. Point Transformer 2021. 10.48550/arXiv.2012.09164.
46. Misra, I.; Girdhar, R.; Joulin, A. An End-to-End Transformer Model for 3D Object Detection.; 2021; pp. 2906–2917.
47. Wang, Z.; Li, Y.; Chen, X.; Zhao, H.; Wang, S. Uni3DETR: Unified 3D Detection Transformer 2023. 10.48550/arXiv.2310.05699.
48. Zhou, Y.; Tuzel, O. VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection 2017. 10.48550/arXiv.1711.06396.
49. Yan, Y.; Mao, Y.; Li, B. SECOND: Sparsely Embedded Convolutional Detection. *Sensors* **2018**, *18*, 3337. 10.3390/s18103337.
50. Wang, B.; An, J.; Cao, J. Voxel-FPN: Multi-Scale Voxel Feature Aggregation in 3D Object Detection from Point Clouds 2019. 10.48550/arXiv.1907.05286.
51. Ye, M.; Xu, S.; Cao, T. HVNet: Hybrid Voxel Network for LiDAR Based 3D Object Detection 2020. 10.48550/arXiv.2003.00186.
52. Hirschmuller, H. Stereo Processing by Semiglobal Matching and Mutual Information. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 328–341. 10.1109/TPAMI.2007.1166.
53. Chen, W.; Shang, G.; Ji, A.; Zhou, C.; Wang, X.; Xu, C.; Li, Z.; Hu, K. An Overview on Visual SLAM: From Tradition to Semantic. *Remote Sens.* **2022**, *14*, 3010. 10.3390/rs14133010.
54. Mayer, N.; Ilg, E.; Häusser, P.; Fischer, P.; Cremers, D.; Dosovitskiy, A.; Brox, T. A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016; pp. 4040–4048. 10.1109/CVPR.2016.438.
55. Saxena, A.; Sun, M.; Ng, A.Y. Make3D: Learning 3D Scene Structure from a Single Still Image. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *31*, 824–840. 10.1109/TPAMI.2008.132.
56. Liu, F.; Shen, C.; Lin, G.; Reid, I. Learning Depth from Single Monocular Images Using Deep Convolutional Neural Fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 2024–2039. 10.1109/TPAMI.2015.2505283.
57. Xie, J.; Girshick, R.; Farhadi, A. Deep3D: Fully Automatic 2D-to-3D Video Conversion with Deep Convolutional Neural Networks 2016. 10.48550/arXiv.1604.03650.
58. Socher, R.; Huval, B.; Bhat, B.; Manning, C.; Ng, A. Convolutional-Recursive Deep Learning for 3D Object Classification. *NIPS* **2012**, *1*.
59. Mousavian, A.; Anguelov, D.; Flynn, J.; Kosecka, J. 3D Bounding Box Estimation Using Deep Learning and Geometry 2017. 10.48550/arXiv.1612.00496.
60. Chen, X.; Kundu, K.; Zhang, Z.; Ma, H.; Fidler, S.; Urtasun, R. Monocular 3D Object Detection for Autonomous Driving.; 2016; pp. 2147–2156.
61. Chen, X.; Ma, H.; Wan, J.; Li, B.; Xia, T. Multi-View 3D Object Detection Network for Autonomous Driving 2017. 10.48550/arXiv.1611.07759.
62. Gupta, S.; Girshick, R.; Arbeláez, P.; Malik, J. Learning Rich Features from RGB-D Images for Object Detection and Segmentation. In Proceedings of the Computer Vision – ECCV 2014; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer International Publishing: Cham, 2014; pp. 345–360. 10.1007/978-3-319-10584-0_23.
63. Luo, Q.; Ma, H.; Tang, L.; Wang, Y.; Xiong, R. 3D-SSD: Learning Hierarchical Features From RGB-D Images for Amodal 3D Object Detection. *Neurocomputing* **2020**, *378*, 364–374. 10.1016/j.neucom.2019.10.025.
64. Qi, C.R.; Liu, W.; Wu, C.; Su, H.; Guibas, L.J. Frustum PointNets for 3D Object Detection from RGB-D Data 2018. 10.48550/arXiv.1711.08488.

65. Wang, Y.; Xu, S.; Zell, A. Real-time 3D Object Detection from Point Clouds using an RGB-D Camera.; 2024; pp. 407–414.
66. Shen, X.; Stamos, I. Frustum VoxNet for 3D object detection from RGB-D or Depth images.; 2020; pp. 1698–1706.
67. Ren, Z.; Sudderth, E.B. Three-Dimensional Object Detection and Layout Prediction Using Clouds of Oriented Gradients.; 2016; pp. 1525–1533.
68. Ren, Z.; Sudderth, E.B. 3D Object Detection with Latent Support Surfaces.; 2018; pp. 937–946.
69. Song, S.; Xiao, J. Deep Sliding Shapes for Amodal 3D Object Detection in RGB-D Images.; 2016; pp. 808–816.
70. Qi, C.R.; Litany, O.; He, K.; Guibas, L.J. Deep Hough Voting for 3D Object Detection in Point Clouds.; 2019; pp. 9277–9286.
71. Hatır, M.E.; İnce, İ.; Korkanç, M. Intelligent Detection of Deterioration In Cultural Stone Heritage. *J. Build. Eng.* **2021**, *44*, 102690. 10.1016/j.jobe.2021.102690.
72. Mishra, M.; Barman, T.; Ramana, G.V. Artificial Intelligence-Based Visual Inspection System for Structural Health Monitoring of Cultural Heritage. *J. Civ. Struct. Health Monit.* **2022**. 10.1007/s13349-022-00643-8.
73. Liu, C.; M.E. Sepasgozar, S.; Shirowzhan, S.; Mohammadi, G. Applications of Object Detection in Modular Construction Based on a Comparative Evaluation of Deep Learning Algorithms. *Constr. Innov.* **2021**, *22*, 141–159. 10.1108/CI-02-2020-0017.
74. Verges-Belmin, V.; Stone (ISCS), I.S.C. For *Illustrated Glossary on Stone Deterioration Patterns*; ICOMOS, 2008; ISBN 978-2-918086-00-0.
75. Kwon, D.; Yu, J. Automatic Damage Detection of Stone Cultural Property Based on Deep Learning Algorithm. *ISPRS - Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2019**, *XLII-2/W15*, 639–643. 10.5194/isprs-archives-XLII-2-W15-639-2019.
76. Wang, N.; Zhao, X.; Zhao, P.; Zhang, Y.; Zou, Z.; Ou, J. Automatic Damage Detection of Historic Masonry Buildings Based on Mobile Deep Learning. *Autom. Constr.* **2019**, *103*, 53–66. 10.1016/j.autcon.2019.03.003.
77. Idjaton, K.; Desquesnes, X.; Treuillet, S.; Brunetaud, X. Transformers with YOLO Network for Damage Detection in Limestone Wall Images. In Proceedings of the Image Analysis and Processing. ICIAP 2022 Workshops; Mazzeo, P.L., Frontoni, E., Sclaroff, S., Distante, C., Eds.; Springer International Publishing: Cham, 2022; pp. 302–313. 10.1007/978-3-031-13324-4_26.
78. Karimi, N.; Valibeig, N.; Rabiee, H.R. Deterioration Detection in Historical Buildings with Different Materials Based on Novel Deep Learning Methods with Focusing on Isfahan Historical Bridges. *Int. J. Archit. Herit.* **2023**, *0*, 1–13. 10.1080/15583058.2023.2201576.
79. Guerrieri, M.; Parla, G. Flexible and Stone Pavements Distress Detection and Measurement by Deep Learning and Low-Cost Detection Devices. *Eng. Fail. Anal.* **2022**, *141*, 106714. 10.1016/j.engfailanal.2022.106714.
80. Tatzel, L.; Tamimi, O.A.; Haueise, T.; Puente León, F. Image-Based Modelling and Visualisation of The Relationship Between Laser-Cut Edge And Process Parameters. *Opt. Laser Technol.* **2021**, *141*, 107028. 10.1016/j.optlastec.2021.107028.
81. Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; Samek, W. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE* **2015**, *10*, e0130140. 10.1371/journal.pone.0130140.
82. Pathak, R.; Saini, A.; Wadhwa, A.; Sharma, H.; Sangwan, D. An Object Detection Approach for Detecting Damages in Heritage Sites Using 3-D Point Clouds And 2-D Visual Data. *J. Cult. Herit.* **2021**, *48*, 74–82. 10.1016/j.culher.2021.01.002.
83. Grilli, E.; Dininno, D.; Petrucci, G.; Remondino, F. From 2D to 3D Supervised Segmentation and Classification for Cultural Heritage Applications. In Proceedings of the The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences; Copernicus GmbH, 2018; Vol. XLII–2, pp. 399–406. 10.5194/isprs-archives-XLII-2-399-2018.
84. Grilli, E.; Remondino, F. Classification of 3D Digital Heritage. *Remote Sens.* **2019**, *11*, 847. 10.3390/rs11070847.
85. Lowe, D.G. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. 10.1023/B:VISI.0000029664.99615.94.
86. Bay, H.; Ess, A.; Tuytelaars, T.; Van Gool, L. Speeded-Up Robust Features (SURF). *Comput. Vis. Image Underst.* **2008**, *110*, 346–359. 10.1016/j.cviu.2007.09.014.
87. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An Efficient Alternative to SIFT Or SURF. In Proceedings of the 2011 International Conference on Computer Vision; 2011; pp. 2564–2571. 10.1109/ICCV.2011.6126544.
88. DeTone, D.; Malisiewicz, T.; Rabinovich, A. SuperPoint: Self-Supervised Interest Point Detection and Description.; 2018; pp. 224–236.
89. Dusmanu, M.; Rocco, I.; Pajdla, T.; Pollefeys, M.; Sivic, J.; Torii, A.; Sattler, T. D2-Net: A Trainable CNN for Joint Description and Detection of Local Features.; 2019; pp. 8092–8101.

90. Revaud, J.; De Souza, C.; Humenberger, M.; Weinzaepfel, P. R2D2: Reliable and Repeatable Detector and Descriptor. In Proceedings of the Advances in Neural Information Processing Systems; Curran Associates, Inc., 2019; Vol. 32.
91. Stathopoulou, E.-K.; Remondino, F. Semantic Photogrammetry – Boosting Image-Based 3d Reconstruction with Semantic Labeling. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2019**, *XLII-2/W9*, 685–690. 10.5194/isprs-archives-XLII-2-W9-685-2019.
92. Stathopoulou, E.-K.; Remondino, F. Multi-View Stereo with Semantic Priors. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2019**, *XLII-2-W15*, 1135–1140. 10.5194/isprs-archives-XLII-2-W15-1135-2019.
93. Stathopoulou, E.K.; Battisti, R.; Cernea, D.; Remondino, F.; Georgopoulos, A. Semantically Derived Geometric Constraints for MVS Reconstruction of Textureless Areas. *Remote Sens.* **2021**, *13*, 1053. 10.3390/rs13061053.
94. Liu, Z.; Brigham, R.; Long, E.R.; Wilson, L.; Frost, A.; Orr, S.A.; Grau-Bové, J. Semantic Segmentation and Photogrammetry of Crowdsourced Images to Monitor Historic Facades. *Herit. Sci.* **2022**, *10*, 27. 10.1186/s40494-022-00664-y.
95. Civera, J.; Gálvez-López, D.; Riazuelo, L.; Tardós, J.D.; Montiel, J.M.M. Towards Semantic SLAM Using a Monocular Camera. In Proceedings of the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems; 2011; pp. 1277–1284. 10.1109/IROS.2011.6094648.
96. Yu, C.; Liu, Z.; Liu, X.-J.; Xie, F.; Yang, Y.; Wei, Q.; Fei, Q. DS-SLAM: A Semantic Visual SLAM towards Dynamic Environments. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2018; pp. 1168–1174. 10.1109/IROS.2018.8593691.
97. Truong, P.H.; You, S.; Ji, S. Object Detection-based Semantic Map Building for A Semantic Visual SLAM System. In Proceedings of the 2020 20th International Conference on Control, Automation and Systems (ICCAS); 2020; pp. 1198–1201. 10.23919/ICCAS50221.2020.9268441.
98. Sun, Z.; Xu, Y.; Hoegner, L.; Stilla, U. Classification of MLS Point Clouds in Urban Scenes Using Detrended Geometric Features From Supervoxel-Based Local Contexts. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2018**, *IV–2*, 271–278. 10.5194/isprs-annals-IV-2-271-2018.
99. Grilli, E.; Özdemir, E.; Remondino, F. Application of Machine and Deep Learning Strategies for The Classification of Heritage Point Clouds. In Proceedings of the The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences; Copernicus GmbH, 2019; Vol. XLII-4-W18, pp. 447–454. 10.5194/isprs-archives-XLII-4-W18-447-2019.
100. Teruggi, S.; Grilli, E.; Russo, M.; Fassi, F.; Remondino, F. A Hierarchical Machine Learning Approach for Multi-Level and Multi-Resolution 3D Point Cloud Classification. *Remote Sens.* **2020**, *12*, 2598. 10.3390/rs12162598.
101. Zhang, K.; Teruggi, S.; Fassi, F. Machine Learning Methods for Unesco Chinese Heritage: Complexity and Comparisons. In Proceedings of the The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences; Copernicus GmbH, 2022; Vol. XLVI-2-W1-2022, pp. 543–550. 10.5194/isprs-archives-XLVI-2-W1-2022-543-2022.
102. Zhang, K.; Teruggi, S.; Ding, Y.; Fassi, F. A Multilevel Multiresolution Machine Learning Classification Approach: A Generalization Test on Chinese Heritage Architecture. *Heritage* **2022**, *5*, 3970–3992. 10.3390/heritage5040204.
103. Grilli, E.; Remondino, F. Machine Learning Generalisation across Different 3D Architectural Heritage. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 379. 10.3390/ijgi9060379.
104. Cao, Y.; Teruggi, S.; Fassi, F.; Scaioni, M. A Comprehensive Understanding of Machine Learning and Deep Learning Methods for 3D Architectural Cultural Heritage Point Cloud Semantic Segmentation. In Proceedings of the Geomatics for Green and Digital Transition; Borgogno-Mondino, E., Zamperlin, P., Eds.; Springer International Publishing: Cham, 2022; pp. 329–341. 10.1007/978-3-031-17439-1_24.
105. Malinverni, E.S.; Pierdicca, R.; Paolanti, M.; Martini, M.; Morbidoni, C.; Matrone, F.; Lingua, A. Deep Learning for Semantic Segmentation of 3D Point Cloud. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2019**, *XLII-2/W15*, 735–742. 10.5194/isprs-archives-XLII-2-W15-735-2019.
106. Sural, S.; Qian, G.; Pramanik, S. Segmentation and Histogram Generation Using the HSV Color Space for Image Retrieval. In Proceedings of the Proceedings. International Conference on Image Processing; 2002; Vol. 2, p. II–II. 10.1109/ICIP.2002.1040019.
107. Pierdicca, R.; Paolanti, M.; Matrone, F.; Martini, M.; Morbidoni, C.; Malinverni, E.S.; Frontoni, E.; Lingua, A.M. Point Cloud Semantic Segmentation Using a Deep Learning Framework for Cultural Heritage. *Remote Sens.* **2020**, *12*, 1005. 10.3390/rs12061005.
108. Matrone, F.; Grilli, E.; Martini, M.; Paolanti, M.; Pierdicca, R.; Remondino, F. Comparing Machine and Deep Learning Methods for Large 3D Heritage Semantic Segmentation. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 535. 10.3390/ijgi9090535.
109. Matrone, F.; Lingua, A.; Pierdicca, R.; Malinverni, E.S.; Paolanti, M.; Grilli, E.; Remondino, F.; Murtiyoso, A.; Landes, T. A Benchmark for Large-Scale Heritage Point Cloud Semantic Segmentation. In Proceedings of the The International Archives of the Photogrammetry, Remote Sensing and Spatial Information

Sciences; Copernicus GmbH, 2020; Vol. XLIII-B2-2020, pp. 1419–1426. 10.5194/isprs-archives-XLIII-B2-2020-1419-2020.
110. Cao, Y.; Scaioni, M. A Pre-Training Method for 3d Building Point Cloud Semantic Segmentation. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2022**, *V-2–2022*, 219–226. 10.5194/isprs-annals-V-2-2022-219-2022.
111. Cao, Y.; Scaioni, M. 3DLEB-Net: Label-Efficient Deep Learning-Based Semantic Segmentation of Building Point Clouds at LoD3 Level. *Appl. Sci.* **2021**, *11*, 8996. 10.3390/app11198996.
112. Cao, Y.; Previtali, M.; Scaioni, M. Understanding 3d Point Cloud Deep Neural Networks by Visualization Techniques. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2020**, *XLIII-B2-2020*, 651–657. 10.5194/isprs-archives-XLIII-B2-2020-651-2020.
113. Cao, Y.; Scaioni, M. Label-Efficient Deep Learning-Based Semantic Segmentation of Building Point Clouds at Lod3 Level. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2021**, *XLIII-B2-2021*, 449–456. 10.5194/isprs-archives-XLIII-B2-2021-449-2021.
114. Armeni, I.; Sax, S.; Zamir, A.R.; Savarese, S. Joint 2D-3D-Semantic Data for Indoor Scene Understanding 2017. 10.48550/arXiv.1702.01105.
115. Ma, J.W.; Czerniawski, T.; Leite, F. Semantic Segmentation of Point Clouds of Building Interiors with Deep Learning: Augmenting Training Datasets with Synthetic BIM-Based Point Clouds. *Autom. Constr.* **2020**, *113*, 103144. 10.1016/j.autcon.2020.103144.
116. Landrieu, L.; Simonovsky, M. Large-Scale Point Cloud Semantic Segmentation with Superpoint Graphs.; 2018; pp. 4558–4567.
117. Li, G.; Muller, M.; Thabet, A.; Ghanem, B. DeepGCNs: Can GCNs Go As Deep As CNNs?; 2019; pp. 9267–9276.
118. Hu, Q.; Yang, B.; Xie, L.; Rosa, S.; Guo, Y.; Wang, Z.; Trigoni, N.; Markham, A. RandLA-Net: Efficient Semantic Segmentation of Large-Scale Point Clouds.; 2020; pp. 11108–11117.