A Saliency Aware CNN-Based 3D Model Simplification and Compression Framework for Remote Inspection of Heritage Sites

STAVROS NOUSIAS\textsuperscript{1,2}, (Member, IEEE), GERASIMOS ARVANITIS\textsuperscript{2}, ARIS S. LALOS\textsuperscript{2}, (Senior Member, IEEE), GEORGE PAVLIDIS\textsuperscript{3}, (Senior Member, IEEE), CHRISTOS KOUALAMAS\textsuperscript{2}, (Senior Member, IEEE), ATHANASIOS KALOGERAS\textsuperscript{2}, (Senior Member, IEEE), AND KONSTANTINOS MOUSTAKAS\textsuperscript{1}, (Senior Member, IEEE)

\textsuperscript{1}Department of Electrical and Computer Engineering, University of Patras, 26504 Patras, Greece
\textsuperscript{2}Industrial Systems Institute, Athena Research Center, 26504 Patras, Greece
\textsuperscript{3}Institute for Language and Speech Processing, Athena Research Center, 15125 Athens, Greece

Corresponding author: Stavros Nousias (snousias@upatras.gr)

This work was supported by the European Union Horizon 2020 Research and Innovation Program “WARMEST -or loW Altitude Remote sensing for the Monitoring of the state of cultural hEritage Sites: building an inTegrated model for maintenance” through the Marie Skłodowska-Curie Grant Agreement 777981.

\textbf{ABSTRACT} Nowadays, the preservation and maintenance of historical objects is the main priority in the area of the heritage culture. The new generation of 3D scanning devices and the new assets of technological improvements have created a fertile ground for developing tools that could facilitate challenging tasks which traditionally required a huge amount of human effort and specialized knowledge of experts (e.g., a detailed inspection of defects in a historical object due to aging). These tasks demand more human effort, especially in some special cases, such as the inspection of a large-scale or remote object (e.g., tall columns, the roof of historical buildings, etc.), where the preserver expert does not have easy access to it. In this work, we propose a saliency aware compression and simplification framework for efficient remote inspection of Structure From Motion (SFM) reconstructed heritage 3D models. More specifically, we present a Convolutional Neural Network (CNN) based saliency map extraction pipeline that highlights the most important information of a 3D model. These include geometric details such as the fine features of the model or surface defects. An extensive experimental study, using a large number of real SFM reconstructed heritage 3D models, verifies the effectiveness and the robustness of the proposed method providing very promising results and draws future directions.

\textbf{INDEX TERMS} Saliency-aware compression and simplification, SFM reconstruction, cultural heritage 3D models.

\textbf{I. INTRODUCTION} Nowadays, thanks to the ease to create digital 3D content, a great amount of information can be easily captured and stored instantly. However, the information, acquired by 3D scanners, is usually immense, creating dense 3D models. Handling the generated geometries efficiently is a challenging task, due to high computational complexity and computational cost. On the other hand, geometric details of a 3D model, such as high-frequency features, and surface areas consisting of points-of-interest (e.g., cracks in an ancient vase) must be preserved, keeping their original high-resolution quality. These salient areas must be handled differently (i.e., anisotropically) from the rest 3D model (i.e., smooth or flat areas), especially in applications that deal with sensitive content such as the 3D representation of historical objects. In this type of application, any perceptual detail or even any possible defect of the object must be preserved since they represent importantly valuable information that can be used through the process of object’s maintenance and preservation.

Defects are referred to as an abnormal condition that affects the original structure of a heritage object (e.g., statue or a building). They may originate from the slow and inevitable ageing processes due to: (1) material deterioration, (2) modifications by humans (e.g., in the urban environment), and
(3) climate and environmental changes [1]. The preservation and maintenance of heritage objects focus on protecting them from being destroyed and trying to preserve both their aesthetic and historical values. Conservation emphasizes the importance of preserving cultural properties and tries to extend the life of historical objects [2].

Monitoring and visualization of historical objects is a necessary process for visual inspection, in order to constantly detect if any intervention of a professional expert is required. Therefore, it seems necessary to develop tools for implementing an effective preventive policy, which will take into account all the conservation requirements. However, visualization is not always enough for the preservation of historic buildings since the conservation professionals do not only need to navigate the 3D model but also to perform spatial and multi-criteria queries in a virtual 3D environment for making decisions. This requirement is even more urgent in cases where the buildings demonstrate some critical evolution, cracks or any potential collapse [3]. Parametric tools can reduce human involvement, providing at the same time reliable inspection results in an automatic process. Such tools can facilitate the work of professionals, highlighting salient region-of-interest and providing them with other suggestions and reports.

The use of building information modelling (BIM) in the case of historical buildings and archaeological infrastructures has engaged the scientific community, over the last few years, to a new trend called historic building information modelling (HBIM). HBIM uses a reverse modelling process that takes place from transforming point clouds into a 3D models [4]. HBIM is considered a promising resource for the planned conservation of historical assets, thanks to its capability of archiving and organizing all the information about a building. However, a considerable amount of information has to be stored and transmitted efficiently and quickly [5], so as to be accessible and modifiable by many different professionals involved in the same project [6]. This need mandates the use of compressed models since the initially scanned 3D models are very dense (with millions of vertices). As a result, conventional software cannot handle them without further processing (e.g., simplification and compression).

The workflow of a preservation process starts with data acquisition, which represents the digitized information of the historical object-of-interest, collected by different sensors or devices. The point cloud is subsequently pre-processed and reduced with various systems and dedicated software. During the point cloud conversion to 3D model, several decisions have to be made, such as the range of precision, also called as Level-of-Detail (LoD) [7]. The Level-of-Detail in historic objects represents the expected complexity, accuracy and change of the small-scaled features (i.e., fine details). Regarding the aspect of spatial information, the 3D historical models are being generated at different levels of detail and scales using methodologies based on different accurate data acquisition techniques [3]. The latest survey techniques for creating 3D heritage objects are able to capture very fine details but, at the same time, the conversion to a 3D model is still intricate and time-consuming. Further research is needed in order to identify which LoD is the ideal for different tasks or to easily create simplified models on demand, especially for scalable coding applications using AR/VR technologies.

Motivated by the aforementioned open issues our approach extracts importance maps from 3D meshes utilizing convolutional neural networks facilitating the utilization of a multiscale feature aware simplification and compression framework. More specifically: i) We have developed novel methods for extracting efficiently, saliency maps with high resolution from large 3D heritage models. These methods are based on deep architectures that also perform localization, i.e., a class label corresponds to each face, has been successfully used to allow training with a small set of segmented SFM reconstructed 3D models. The ground truth data were generated using traditional mesh denoising and segmentation approaches. ii) We have proposed novel approaches for simplification and compression that efficiently exploit the geometric features at multiple levels of granularity, extracted from the aforementioned novel saliency mapping method, allowing the preservation of defects even in extremely high simplification and compression ratios. iii) We present an extensive evaluation of the presented approaches using dense SFM 3D models reconstructed from drone imagery taken at the sites of Patio de Leones in Granada and Santa Croce in Florence. The rest of this paper is organized as follows: Section II presents the related work. Section III analyzes the extraction of saliency maps with sparse and deep priors. Additionally, for the sake of clarity Table I presents notation and definitions. Section IV presents the experimental evaluation and the results of our approach, while Section V concludes this work.

II. RELATED WORK

A. SALIENCY MAP EXTRACTION

The extraction of visual saliency has been previously investigated by numerous researchers. Nevertheless, it continues to be a challenging task due to the abstract meaning of saliency. Saliency extraction represents a stimulus-driven activation process [8] where the salient part of the scene differs from its neighboring region [9] due to lack of correlation [10]. Psychophysical studies have shown that low-level salient features are processed in the early visual cortex and extract the most important information of the visual scene [11]. Saliency is mainly affected by feature type, and contrast, task difficulty, center bias, temporal variations of fixations as a recent study reports [12].

A lot of work, related to salient map extraction, has occurred in the area of image [13]–[17] and video processing [18]–[20]. Recently, this trend has started to attract the attention of the researchers in the area of geometry processing, where the saliency has been defined as a measure of regional importance, showing its benefits in many applications and opening a new road of processing possibilities.
TABLE 1. Notations and definitions.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A/A/\alpha/\alpha$</td>
<td>Tensor / Matrix / Vector / Scalar</td>
</tr>
<tr>
<td>$I_n$</td>
<td>Identity matrix</td>
</tr>
<tr>
<td>$\mathbb{R}$</td>
<td>Field of real numbers</td>
</tr>
<tr>
<td>$A_{i,j,k}$</td>
<td>$(i,j,k)$th element of $A$</td>
</tr>
<tr>
<td>$A_{i,j}(i,j,i,j)$</td>
<td>slice of $A$</td>
</tr>
<tr>
<td>$n$</td>
<td>number of vertices</td>
</tr>
<tr>
<td>$n_f$</td>
<td>number of faces</td>
</tr>
<tr>
<td>$n_p$</td>
<td>number of normal vectors in a single patch</td>
</tr>
<tr>
<td>$V$</td>
<td>Vertices matrix</td>
</tr>
<tr>
<td>$v_i$</td>
<td>$i^{th}$ vertex coordinates</td>
</tr>
<tr>
<td>$c_i$</td>
<td>$i^{th}$ centroid coordinates</td>
</tr>
<tr>
<td>$n_i$</td>
<td>coordinates for $i^{th}$ vertex normal</td>
</tr>
<tr>
<td>$n_e$</td>
<td>coordinates for $i^{th}$ centroid normal</td>
</tr>
<tr>
<td>$n_{f,i}$</td>
<td>$i^{th}$ face normal</td>
</tr>
<tr>
<td>$\mathbb{R}$</td>
<td>Covariance matrix</td>
</tr>
</tbody>
</table>
| $P_t$    | Patch matrix containing normal coordinates in the
|          | spectral saliency case                           |
| $P_t$    | Patch tensor containing normal coordinates in the
|          | CNN case                                         |
| $M_L$    | Low rank matrix                                  |
| $M_S$    | Sparse matrix                                     |
| $L$      | Laplacian matrix                                 |
| $A$      | Adjacency matrix                                 |
| $D$      | Diagonal matrix                                  |
| $T$      | Interval $[0,1]$                                 |
| $Q$      | Interval $[0,1]^p$                               |
| $\delta_i$ | Delta coordinate of the $i^{th}$ vertex          |
| $Q(\cdot)$ | Quantization function                            |
| $1$      | Vector with coordinates $[1,1,1]$                |

3D mesh saliency was coined, as a term, by Lee et al. [21],
that employed Gaussian weighted-averaged mean curvature-

based approach to compute a perceptual metric. Several other
approaches followed, utilizing spectral methods [22], [23],
curvature-based methods [21], [24] multiscale descriptors
[25], entropy-based methods [26] and hybrid methods [27] taking into account both color and color.

Wu et al. presented local multiscale descriptors with global rarity [25], using rotationally invariant metrics to capture local features and dissimilarity based estimators to assess global characteristics. Spectral methods for 3D meshes processing [23] were employed by Song and Liu utilizing the spectral attributes of the log-Laplacian formulation. Wei et al. [24] presented a 3D saliency mapping mechanism using the curvature co-occurrence histogram. Nouri et al. [28] proposed a saliency-based metric for the evaluation of the quality between an original and a distorted 3D mesh comparing the structural information. They also [29] proposed a saliency method, using a local vertex descriptor that is used as a basis for similarity measurement and integrated into weighted multiscale saliency features. Zhao et al. [30] proposed a saliency detection method by diffusing a shape index field with a non-local means filter. Their algorithm generates a random center-surround operator to create a saliency map and use the Retinex theory to improve it. Tao et al. [26] proposed an entropy-based saliency approach using the entropy of the normals to depict the local changes in a region. Lau et al. [31] explored the problem of tactile mesh saliency, searching for the parts of a mesh that are more likely to be grasped, pressed, or touched by humans in the real world. Song et al. [22] proposed a method that incorporates global considerations by making use of spectral attributes. An et al. [27] proposed a hybrid saliency taking into account both color and geometric information.

A new direction for the salient map extraction has started to appear by using Convolutional neural networks (CNNs). CNNs have been extensively used for the extraction of saliency maps in image processing as a recent survey reports [32].

A recent study employs CNNs to extract saliency maps on 3D meshes utilizing a multi-view setup [33]. Nousias et al. [34] employed a CNN based mesh saliency extraction approach, utilizing a 3D geometric patch descriptor to classify faces into four classes. Greater class indices correspond to higher saliency values. Yet, our approach generates saliency values in a continuous range focusing on a regression problem instead of a classification. All these CNN based approaches provide very promising results and have the clear benefit of low computational complexity in comparison with the other conventional methods, making them ideal for utilization in some very challenging situations like the processing of very dense models.

B. 3D RECONSTRUCTION

Recently, significant advancements were achieved in structure from motion techniques which, basically stemmed from the integration of the traditional photogrammetry and multiple-view stereo reconstruction. In the structure from motion approach (SfM), an object is photographed from many viewing positions with large overlaps, creating a redundant dataset that serves as input to the solution of the 3D reconstruction inverse problem. Essentially, SfM is a triangulation-based method that operates on large sets of photos of static scenes of rigid bodies, taken from various viewpoints with cameras of unknown characteristics and settings. The method simultaneously estimates the location of surface points (structure) and the pose of the camera (motion). SfM is one of the most studied approaches for image-based 3D reconstruction for more than 30 years in the computer vision community, with research efforts devoted to it being the most intense by any area of computer vision [35].

At the heart of SfM is the solution to the point correspondence problem in a multiple view setting, which includes the detection of image features, matching of the image features across photos, the creation of tracks from matches and the solution of the SfM problem from those tracks. In the SfM context, “tracks” are defined as 3D coordinates of reconstructed points accompanied by a list of the corresponding 2D coordinates in the photos [36]. SfM has been significantly automated by the adoption of advanced feature extraction approaches such as SIFT [37] and SURF [38].

The original version of SfM produces sparse point clouds from point correspondences and is therefore of limited practical value, as it leads to very low-resolution 3D models. The popularity and applicability of SfM have been
significantly boosted when it was successfully combined with Multi-View Stereo (MVS) methods that may result in dense point clouds. In this hybrid now method, “traditional” SfM is being used to estimate the camera parameters and pose that initialises MVS, which takes over to produce the final dense point cloud of the measured objects. A comparison and evaluation of multi-view stereo reconstruction algorithms can be found in [39]. The hybrid SfM-MVS is based on a set of complex and computationally demanding, memory-hungry algorithms. The method is robust in small changes in light and intensity of the colours in the photos, and this makes it even more user-friendly for outdoor or indoor applications.

As with any other 3D reconstruction method, SfM-MVS has its limitations. Since it is based on point correspondences on object surfaces captured in photos, the quality of the 3D reconstruction is largely influenced by the existence of intense morphological features. Applying the method to featureless surfaces or surfaces with significant self-similarity can lead to incorrect point matches among photos and, ultimately, to the introduction of intense noise in the reconstructed geometry, or even to the reconstruction with severe deformations. Several studies have focused on this problem. Research has already focused on the issues of featureless and specular surfaces, the image pre-processing for SfM-MVS optimisation and the selection of the proper acquisition setup [40]–[49].

SfM-MVS has been very successful and tends to become the most widespread and common photogrammetric approach adopted by experts in various domains. When properly applied, it can be of low cost with high quality (in terms of accuracy and resolution) 3D reconstruction results. In particular, the creation of 3D digital replicas of cultural objects by SfM-MVS tends to become the most popular solution, surpassing laser scanning and structured light approaches. The 3D data resulting from SfM-MVS systems are of high quality, and such systems are currently being used in demanding cultural digitisation applications. As SfM-MVS solutions are of relatively low cost and can be automated and easily implemented by non-specialists, they are increasingly becoming more appealing. The popularity of SfM-MVS gets even higher if we also consider the increasing quality and capabilities of modern, low-cost capturing equipment, such as low-end drones, especially after the introduction of less restrictive regulatory frameworks for the usage of these devices [50], [51].

III. LOW LEVEL GEOMETRY PROCESSING USING SPARSE AND DEEP PRIORS
This section presents the pipeline of the proposed approach for the extraction of the 3D saliency mapping. We generated training data from meshes using baseline saliency maps based on traditional methods [52]. After the training process, the generated output can be used for the automatic extraction of saliency mapping for any other new 3D model. Figure 4 briefly presents the pipeline of our approach. We start by separating the whole mesh into \( n_f \) (i.e., equal to the number of centroids) overlapped and equal-sized patches. Then, we follow two different steps for the estimation of the spectral and geometrical saliency. The final result is a combination of these two values. Once the saliency mapping has been estimated, we use it as input for the training of the CNN. We also present some preliminaries and discussion related to the two proposed tools (i.e., saliency-based compression...
and simplification) that can be used for the facilitation of the remote inspection. More specifically, compression facilitates the storage of dense 3D models, also providing an efficient reconstruction quality through the receiver phase. On the other hand, simplification creates a perceptually inferior representation of a model, but it emphasizes real-time transmission. The choice for using compression or simplification mainly depends on the essential requirement of the application (quality of the reconstructed model, the time efficiency of the transmission).

A. Saliency Feature Extraction Using Sparse Modeling

In this paragraph, we will present the conventional methods that are used for the extraction of the ground truth saliency mapping, which will be utilized for the training process of the proposed CNN.

We assume that the reconstructed heritage objects are represented by triangular meshes $\mathcal{M}$ with $n$ vertices $v_i$ and $n_f$ faces $f$. Each vertex $v_i$ is denoted by $v_i = [x_i, y_i, z_i]^T$, $\forall i = 1, \ldots, n$. Each $f_j$ face is a triangle that can be described by its centroid $c_j = (v_{j1} + v_{j2} + v_{j3})/3$ and its outward unit normal: $n_{c_i} = (v_{j2} - v_{j1}) \times (v_{j3} - v_{j1}) / \| (v_{j2} - v_{j1}) \times (v_{j3} - v_{j1}) \|$. Where $v_{j1}, v_{j2}$ and $v_{j3}$ are the position of the vertices that define face $f_j = \{v_{j1}, v_{j2}, v_{j3}\}$, $\forall j = 1, \ldots, n_f$.

1) Spectral Saliency Estimation

For each face $f_i$ of the mesh, we estimate a corresponding patch $P_i \in \mathbb{R}^{(k-1) \times 3}$, consisting of the $k - 1$ neighboring centroid normals: $P_i = [n_{c_i}, n_{c_i}, n_{c_i}, \ldots, n_{c_i}]^T$ $\forall i = 1, \ldots, n_f$. This matrix $P_i$ is used for the estimation of the covariance matrices $R_i = P_i^T P_i \in \mathbb{R}^{3 \times 3}$, and a diagonal matrix $\Lambda = \text{diag}(\lambda_{i1}, \lambda_{i2}, \lambda_{i3})$, consisting of the corresponding eigenvalues $\lambda_{ij}$, $\forall j = 1, 2, 3$. Finally, the spectral saliency $s_{iL}$ of a centroid $c_i$ is denoted as the value given by the inverse $l^2$-norm of the corresponding eigenvalues:

$$s_{iL} = \frac{1}{\sqrt{\lambda_{i1}^2 + \lambda_{i2}^2 + \lambda_{i3}^2}} \quad \forall i = 1, \ldots, n_f \quad (1)$$

As has been discussed in [52], if the value of the $l^2$-norm of the eigenvalues (i.e., $\sqrt{\lambda_{i1}^2 + \lambda_{i2}^2 + \lambda_{i3}^2}$) is high, then it indicates that the centroid lies in a flat area and the corresponding saliency value is small since the centroid is not characterized as a significantly important feature. On the other hand, if the value of the $l^2$-norm is small, then it results in large saliency values, characterizing the specific centroid (i.e., feature) as perceptually important (e.g. corner or edge). The aforementioned behavior of the $l^2$-norm can be easily justified by the fact that the centroid normal of a face lying in a flat area is represented by one dominant eigenvector, the corresponding the eigenvalue of which has a huge value. On the other hand, the centroid normal of a face, lying in a corner, is represented by three eigenvectors that correspond to eigenvalues with small but almost equal amplitude [53].

2) Geometrical Saliency Estimation

To estimate the geometrical saliency, we apply a Robust Principal Component Analysis (RPCA) approach to the three matrices $E_l \in \mathbb{R}^{n_f \times (k+1)}$ $\forall l \in \{x, y, z\}$ containing the centroid normals of the corresponding coordinates, as described in [52]. To reconstruct these matrices, we use the $P_i$ matrices according to: $E_l = \{P_{l1}; P_{l2}; \ldots; P_{ln_f}\}$, taking advantage of the geometrical coherence between neighbouring centroid normals.

For the decomposition of a matrix $E$ into a low-rank matrix $M_L$ and a sparse matrix $M_S$ by solving the following equation:

$$\arg \min_{M_L, M_S} \| M_L \|_\ast + \lambda \| M_S \|_1, \quad \text{s.t. } M_L + M_S = E \quad (2)$$

where $\| M_L \|_\ast$ denotes the nuclear norm of a matrix $M_L$. By the decomposition, the low-rank $M_{Lx}, M_{Ly}, M_{Lz}$ and
sparse $M_S$, $M_S$, $M_S$ matrices are estimated. Nevertheless, for the estimation of the geometric saliency feature $s_2$, only the values of the first column of the sparse matrices is needed, according to:

$$s_{2i} = \sqrt{s_{i1}^2 + s_{i2}^2 + s_{i3}^2} \quad \forall i = 1, \ldots, n_f$$

(3)

where $s_{1i}$ denotes the scalar value of the $i^{th}$ row, of the $1^{st}$ column, of the $S$ matrix. Both the values of the spectral and geometrical saliency are normalized between [0-1], according to:

$$\tilde{s}_{xi} = \frac{s_{xi} - \min(s_{xi})}{\max(s_{xi}) - \min(s_{xi})} \quad \forall i = 1, \ldots, n_f, x = \{1, 2\}$$

(4)

The centroid saliency is estimated as the mean value of the corresponding normalized spectral $\tilde{s}_1$ and geometrical $\tilde{s}_2$ values, according to:

$$s_{ci} = \frac{\tilde{s}_{1i} + \tilde{s}_{2i}}{2} \quad \forall i = 1, \ldots, n_f$$

(5)

and finally, the vertex saliency is estimated according to:

$$s_i = \frac{\sum_{j \in N_i} s_{cj}}{|N_i|} \quad \forall i = 1, \ldots, n$$

(6)

B. SALIENCY FEATURE EXTRACTION USING CONVOLUTIONAL NEURAL NETWORKS

To generate baseline saliency maps we employ the approach presented in Section III-A. A geometric descriptor operating
in a sliding patch setup formulates the input for the deep network. The patch descriptor allows for efficient training with relatively small dataset. The geometric patch descriptor characterizes each face with the normal vector distribution of its neighbouring faces and assigns to each patch an estimated groundtruth saliency value. The patch descriptor ensures scale, translation and rotation invariant representation for the face under examination facilitating efficient training. Employing normals instead of positions allows for scale and rotation invariance. Furthermore, using face normals instead of vertex normals allows for higher coherency since each face shares a common edge with three neighbours but each vertex shares a common edge with a varying number of neighbours.

We assume that \( N_{pi} \) is the set of indices of the \( k \) neighbouring faces of face \( i \). \( N_{pi} \) is sorted using the successive rings rule depicted in Figure 5. Matrix \( P_i \in \mathbb{R}^{3 \times (k-1)} \) contains all normal coordinates in set \( N_{pi} \). Furthermore, to achieve rotation invariance we assume a local coordinate system and rotate the patch by angle \( \delta_n \) around rotation axis \( a_n \) so that

\[
\frac{1}{N_p} \sum_{j \in N_{pi}} A_j n_j = n_{const} \tag{7}
\]

where \( A_j \) is the triangle area and \( n_{const} \) is an arbitrary vector. The rotation allows the patch to be rotation-invariant across local x and y axes, and the face arrangement itself renders the patch descriptor rotation invariant with respect to local z-axis.

Furthermore, \( P_i \) is reshaped into patch \( P_i' \in \mathbb{R}^{3 \times w \times w} \) corresponding to face \( f_i \) and containing \( N_{p} = w^2 \) ordered neighbouring face normals \( n_j \equiv P_{j+k,l} \), where \( w \) is set to \( w = 32, j = H(k, l) \) and \( H \) is a space filling curve [54].

Finally, since the values of the normal coordinates in \( P_i' \) range in \([-1, 1]\) they are transformed to \([0, 1]\) with the following expression.

\[
P_i' = \frac{P_i + 1}{2} \tag{8}
\]

For the arrangement of face normals \( n_j \) within patch \( P_i' \) we employ space-filling curves (SFCs) [55]. SFCs are curves traversing all points of the n-dimensional space, thus inducing an order on those points. SFCs are bijective functions so that

\[
f : [0, 1] \rightarrow [0, 1]^2 \tag{9}
\]

A widely used SFC curve is the classical Hilbert space-filling curve, visualized in Figure 10. Hilbert space-filling curve is defined by the partitioning of the interval \( I = [0, 1] \) into \( 4^n \) subintervals of length \( 4^{-n} \) and the square \( Q = [0, 1]^2 \) into \( 4^n \) subsquares of side \( 2^{-n} \) and can be generated by recursion. A bidirectional mapping is generated between the sub-intervals of \( I \) and the sub-squares of \( Q \). Hilbert curve exhibits the highest level of coherency described by the metric defined as the distance along the curve with respect to multi-dimensional Euclidean metric [56]. Hilbert curve is superior to other curves in this respect [57]. Figure 9 depicts different space-filling curves as presented in [54]. In this work, two types of space-filling curves are investigated, simple reshape function, also referred to as sweep space-filling curve and Hilbert space-filling curve. Hilbert curve is employed to improve the local coherency of the patch, allowing the convolutional kernels to capture the relation between locally consistent faces. Figures 6 to 8 highlight the different regions covered by a moving convolutional kernel transversing a patch, ordered either by a Hilbert space-filling curve setting or a sweep space-filling curve setting. Finally, to form the training set, we extract patches from 3D reconstructed meshed and assign saliency.
value to each patch, employing the saliency map extraction process presented in subsection III-A.

Finally, the CNN architecture consists of three convolutional layers and two fully connected layers. Each convolutional layer is succeeded by a max-pooling layer and rectified linear units (ReLUs). The output of the convolutional part is flattened and used as input for the two fully connected layers. The CNN architecture is depicted in Figure 11.

C. SIMPLIFICATION

The reconstructed 3D models, which represent real objects, may be very dense, consisting of millions of vertices. This highly dense information is usually overwhelming, and the original 3D models can be efficiently simplified, keeping only the most representative details, and removing the least important information representing vertices lying in flat or smoothed areas.

Simplification is a low-level application that focuses on representing an object using a lower resolution mesh without errors or with errors that cannot be easily perceived. The main objective of a successful simplification approach is to remove only those vertices which do not offer significant geometric information to the simplified 3D object, and their removal will not change the shape or perceptual details of the 3D object significantly. Most importantly, in the area of heritage maintenance any perceptual detail, both for features and defects, must be preserved.

Following this line of thought, we suggest removing the least perceptually important vertices, preserving only the most salient vertices for the reconstruction of the new simplified 3D model. Nevertheless, to avoid the loss of information on areas with low salient vertices, we suggest the selection of different portions for vertices belonging in different salient
classes. Algorithm 1 briefly describes the step that we follow for the proposed multiscale feature aware simplification.

D. COMPRESSION

3D mesh compression of is another low-level geometry processing application allowing to reconstruct a mesh given the delta coordinates and a subset of mesh vertices according to the approach presented in [58]. In the proposed 3D mesh compression scheme, we are able to encode geometry information allowing the generation of a sequence of levels of detail. Geometric features like high curvature regions, that convey important visual information, bare a high saliency value. Initially, we define the Laplacian matrix

\[ L = D - A \]  \hspace{1cm} (10)

where \( A \) is the adjacency matrix,

\[ A_{ij} = \begin{cases} 1, & \text{if } i, j \in E \\ 0, & \text{otherwise} \end{cases} \]  \hspace{1cm} (11)

and \( D \) is the diagonal matrix, with \( D_{ii} = \sum_{j=1}^{n} A_{ij} \). The reconstruction of the 3D mesh vertices is performed by solving the following sparse linear system

\[
\begin{bmatrix} L & \mathbf{I} \\ \mathbf{I} & V \end{bmatrix} \begin{bmatrix} \mathbf{V} \\ \delta \end{bmatrix} = \mathbf{0}
\]  \hspace{1cm} (12)

where \( \delta = [\delta_1, \delta_2, \ldots, \delta_n] \) and \( \delta_i \) is the delta coordinate of the \( i \)th vertex

\[ \delta_i = \begin{cases} Q(\delta_i), & \text{if } v_i \text{ is a feature} \\ 0, & \text{otherwise} \end{cases} \]  \hspace{1cm} (13)

and \( Q(\cdot) \) is a scalar quantization function. \( \mathbf{V}_a \) is the matrix of uniformly distributed anchor points defined as \( \mathbf{V}_a = \mathbf{I}_i \mathbf{V} \). The anchor points \( \mathbf{V}_a \) are sampled uniformly and quantized to 12 bits. For the delta coordinates we utilize saliency maps to quantize with 12 bits the geometric features while setting the rest of the vertices to zero. The multiscale feature aware compression strategy implies that the resulting saliency maps are separated into \( N \) salient classes. Subsequently, a different ratio is selected from each class. Algorithm 2 presents briefly the multiscale feature aware compression strategy.

IV. SIMULATION STUDIES

In this section, we describe in detail the model acquisition and the CNN training processes. Furthermore, we recommend on saliency extraction results using RPCA-based saliency [58] maps as a baseline and generate corresponding confusion matrices. We also examine simplification and compression cases taking advantage of the vertex importance maps generated by the extracted saliency, meaning that vertices with higher saliency value are treated with greater priority. To compare the reconstruction results with

Algorithm 1 Multiscale Feature Aware Simplification

**Input**: 3D model \( \mathcal{M} \), desired simplification percentage, \( N \) number of classes, saliency map of vertices

**Output**: Simplified model \( \mathcal{M} \)

**// Vertex simplification**

1. Sort the vertices based on the saliency values of the saliency map provided by the CNN approach;
2. Separate the vertices into \( N \) salient classes based on their salient values;
3. Depending on the simplification scenario, different percentages of vertices are selected from each class, giving emphasis to the perceptually salient vertices;
4. The final number of the remaining vertices must satisfy the initial percentage scenario;

**// Connectivity simplification**

5. for each \( i \) removed vertex do
6. Find the nearest vertex that remains;
7. Replace any edge connection of the \( i \)th vertex with the nearest remaining vertex;
8. end
We employed SFM reconstructed 3D models of pillars located in Patio de los Leones and Basilica di Santa Croce. We captured high resolution images from two heritage sites, Patio de los Leones and Basilica di Santa Croce, Florence, Italy. Figure 1 shows a small part of the set of images corresponding to one of the columns of the Brunelleschi’s cloister in Santa Croce, Florence, Italy. Figure 2 presents a class C0 drone, as the miniature device weighs approximately 80gr, it allowed for the safe and automated collection of 5 Mpixel / 1.98mm FL / 80.6° FOV shots from distances around 50cm from the column, as shown in Figure 3, leading to an exploitable capacity of millimeter resolution at the final 3D reconstructed model. However, we need to highlight that due to the non-perfect flight stabilization, multiple photographs had to be captured from each selected viewpoint, in order to choose one with appropriate quality. We employ SFM reconstructed 3D models of pillars located in Patio de los Leones site. They will be referred to as “pillar 14” presented in Figure 13, “pillar 89” presented in 14 and “pillar 90” presented in 15. We also utilized publicly available heritage models, such as “Tycho”, presented in the second row of Figure 13, “terracotta” presented in the third row of Figure 13 and “gladiator” presented in Figure 21.

B. TRAINING

For the training of the CNN a scanned “armchair” 3D model, presented in Figure 4, was employed consisting of 145934 vertices and 291864 faces, generating an equal number of faces, number of training examples. Each training example consists of a tensor \( M \in R^{32 \times 32 \times b \times c} \) and a saliency value \( s \in [0, 1] \) derived from the approach described in subsection IV-A. \( b \) is the batch size, set to \( b = 600 \), and \( c \) the number of channels \( c = 3 \).

For the training of the convolutional neural network, a mean square error cost function was employed with a learning rate of \( 10^{-4} \). The training of the CNN took place in an NVIDIA GeForce GTX 1080 graphics card with 8GB VRAM and compute capability 6.1.

C. SALIENCY EXTRACTION

Figures 12 to 15 present a qualitative evaluation of the extracted saliency maps. Figure 13 presents a qualitative comparison of the extracted saliency maps for the CNN-based approaches with respect to baseline saliency maps. Red colors correspond to higher saliency values while blue colors correspond to lower saliency values. From a qualitative aspect, CNN based saliency maps capture salient features as well as sparse modelling based approach. Furthermore, Hilbert and sweep curve arrangements seem not to differ visually, as a visual comparison of third and fourth columns of Figure 13 reveals. Figure 12 presents a saliency map detail for “pillar14” model. Close inspection reveals that hard corners and features are successfully colored red and flat areas are colored blue. Figure 14 presents saliency map results for “pillar89” model while Figure 15 presents saliency map results for “pillar90” model.
FIGURE 13. Saliency extraction for RPCA-based and CNN-based approaches. (a) Original models (b) RPCA-based saliency extraction (c) CNN based saliency extraction for sweep space filling curve (d) CNN based saliency extraction for Hilbert space filling curve.

FIGURE 14. CNN based saliency map extraction with Hilbert SFC for pillar 89.

To quantitatively measure the CNN prediction with respect to baseline values, we classified the predicted and groundtruth saliency values into 4, 16 and 64 classes and constructed confusion matrices presented in Figure 16. Classes with higher label encompass faces with higher saliency values, i.e. in subfigures, 16a and 16d, class “0”
includes mainly flat or smooth areas, while class “3” mainly sharp angles and curves. Quantitative evaluation reveals that Hilbert curve arrangement captures better curves and features with higher saliency values since sweep curve space-filling approach demonstrates lower accuracy for class 3 and class 2 for subfigures 16a and 16d.
TABLE 2. Execution times and memory requirements for various models as a function of the number of faces.

<table>
<thead>
<tr>
<th>Model</th>
<th>$N_f$</th>
<th>RPCA</th>
<th>CNN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I/O</td>
<td>Inference</td>
<td>Total time</td>
</tr>
<tr>
<td>Gladiator</td>
<td>49994</td>
<td>80</td>
<td>5048</td>
</tr>
<tr>
<td>Terracota</td>
<td>560712</td>
<td>77</td>
<td>6508</td>
</tr>
<tr>
<td>Tycho</td>
<td>544859</td>
<td>88</td>
<td>6485</td>
</tr>
<tr>
<td>Pillar 14</td>
<td>757800</td>
<td>113</td>
<td>8097</td>
</tr>
</tbody>
</table>

TABLE 3. Execution times and memory requirements for "Pillar 14" model as a function of patch size.

<table>
<thead>
<tr>
<th>Model</th>
<th>$N_f$</th>
<th>RPCA</th>
<th>CNN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I/O</td>
<td>Inference</td>
<td>Total Time</td>
</tr>
<tr>
<td>Pillar 14</td>
<td>757800</td>
<td>102</td>
<td>1460</td>
</tr>
<tr>
<td>Pillar 14</td>
<td>757800</td>
<td>113</td>
<td>8097</td>
</tr>
<tr>
<td>Pillar 14</td>
<td>757800</td>
<td>75</td>
<td>17719</td>
</tr>
</tbody>
</table>

FIGURE 17. Execution times and memory requirements as a function of total number of model faces (a) and (c) and patch size (b) and (d).

Quantitative and qualitative evaluation results allow us to conclude that the presented CNN based approaches can be adequately employed to extract saliency maps.

D. PERFORMANCE EVALUATION

Tables 2 and 3 and Figure 17 present the performance evaluation of CNN-based and RPCA-based saliency map extraction approaches in terms of execution times and memory requirements. For the inference part, the CNN based approach demonstrates much lower execution times that the equivalent part of the RPCA based approach. Likewise, minimum memory requirements for CNN are much lower than RPCA-based approach since the RPCA has to be applied to a matrix with memory requirements $N_f \times p \times 3 \times 4$ bytes where $p$ is the patch size. On the other hand, the minimum memory requirement for the CNN is equal to the memory required to store the network parameters $M_N = 70,005, 104$ bytes, and the memory to store the patch, found equal to $M_N + (p \times 3 \times 4 \text{ bytes})$. At that point, one would also notice that above a certain number of model faces or patch size, the computation of RPCA becomes intractable unless a mesh partitioning approach is employed.

E. USE CASES

Simplification and compression are mesh operations performed using rules that signify the importance of specific nodes. Such rules facilitate the selection of certain nodes or certain properties of them such as position to be given higher priority when removing nodes, in the case of simplification, or changing nodes position in the case of compression. Utilizing RPCA or equivalent CNN based approaches, to extract saliency maps, generates importance maps allowing their outcome to be used for both cases.
The simplification process in the field of heritage culture preservation requires a different type of handling with respect to the conventional process of simplification in other domains. More specifically, we do not aim to just create a simplified model, but also to emphasize some specific details with significant meaning such as the geometric features of the model as well as the defected areas. Thanks to our approach, the salient areas are identified and highlighted with higher values. In Figures 20 and 21, we present examples of simplified reconstructed results of different heritage 3D models under different simplification scenarios from a simplification range 30% to 95% percentage. To note here that the simplification percentage denotes the number of vertices that have to be removed. More specifically, in a hypothetical scenario where we apply to a model a simplification ratio equal to 90%, it means that 90% of its initial vertices should be removed and only 10% of them will be used for the reconstruction. As we can observe, the simplified models have been reconstructed concerning the most salient vertices which represent high-frequency features (i.e., perception details and defects). More specifically, the presented simplified models in Figure 20 seem to preserve the defect information even in very high percentages of simplification.

In Tables 4 and 5, we present in details the percentages of vertices that we preserve per each class \( \{c_1, c_2, c_3, c_4\} \) and per each simplification scenario \( \{30\%, 50\%, 70\%, 90\%\} \), where \( c_i \forall i = 1, \ldots, 4 \) represents the \( i \)-th class. The vertices of a higher class are more important than the vertices of a lower class. This is the reason why we keep a higher percentage of vertices from higher classes.

Figure 22 presents the simplification results for two heritage models (i.e., gladiator and pillar14). For an easier
The results showed that the traditional approach of the simplification does not handle differently the perceptually salient vertices (Fig. 22-a) in contrast to our approach that gives emphasis on the preservation of these vertices. The percentages of vertices that we preserve per each class, for the two different approaches (i.e., 4 classes, 8 classes) are presented in Table 6.

Compression of 3D meshes is described in subsection III-D. To make use of the CNN-based saliency map, we assign a saliency value to each vertex.

\[
s_{v_i} = \max\{s_{f_j} : \forall f_j, v_i \in f_j\}
\]  

(14)
A portion of vertices equal to 10%, is sampled uniformly to be used as anchor points $v_a$ and quantized with 12 bit. A sampling of delta coordinates is performed based on saliency value. Delta coordinates of selected vertices are quantized with 12 bit, while the rest are set to zero. The threshold is defined to keep a portion of delta coordinates ranging from 95% to 1% as Figure 18 depicts. For a more comprehensive comparison we also employ the fusion based approach presented in Section III-A, the O3DGC encoder [60] and uniform high pass quantization [61]. For the evaluation of
the efficiency of the compression achieved, we employ the mean theta error metric defined as the mean value of $\theta$ angle difference between the groundtruth face normal position and the reconstructed. Figures 18 and 19 present a comparison of the aforementioned approaches and visualize the error for each vertex. Blue colors correspond to low $\theta$ error, while red colors to high $\theta$ error.

To study the effect of granularity in compression and simplification cases, we compared the simplified and reconstructed models in the case of 8 classes, 4 classes and saliency agnostic uniform sampling. Tables 4, 5 and 6 record the portion of vertices kept per class during the simplification of “gladiator” and “pillar14” models. As Figures 22 and 23 visualize the area around defects is better reconstructed when employing more classes and using an anisotropic strategy for the selection of vertices per class as Figures 22c, 23e and 23f reveal.

V. DISCUSSION

In this work, we presented a fast and effective CNN-based approach for the estimation of saliency maps of large cultural heritage 3D models, consisting of millions of vertices, that represent real historical objects reconstructed using the SFM method.

Our approach aims to highlight and preserve the geometrical salient and perceptually important features of dense models (feature aware), providing also the capability of scalable coding (multiscale).
The conventional methods for saliency map extraction are time-consuming, mainly when they are applied in large models. Data-driven approaches overcome this limitation since they provide both fast and reliable results. One significant advantage of the proposed CNN method is the fact that it does not require a massive dataset for the training process since only one general model can provide millions of coherent instances for efficient learning, due to a pre-processing step that rotates all training data into the same direction.

Furthermore, the outcome of the suggested CNN method is a continuous value, in contrast to previous CNN approaches that only categorized each vertex into a pre-defined salient category. This extension is very beneficial, especially if an adaptive multiscale approach is required, such as in cases where the number of salient classes has to be changed after the training process. Despite the very good results of our method, it can be improved more since we do not use an optimized way for the selection of the vertices per each class. As a future extension of our work, we will search for an automatic way to estimate the number of the preserved vertices per each class, taking into account the minimization of the reconstruction error. Finally, it is essential to mention here that the pipeline as mentioned earlier can also be utilized in other domains that have similar requirements too (e.g., entertainment, military, communication).

REFERENCES


S. Nousias et al.: Saliency Aware CNN-Based 3D Model Simplification and Compression Framework for Remote Inspection
GEORGE PAULIDIS (Senior Member, IEEE) received the Ph.D. degree in electrical engineering, received the Ericsson Awards of Excellence in Telecommunications. He has been working for many research and development projects with emphasis on multimedia systems in culture and education. In 2002, he joined the Athena Research Center, where he is currently the Research Director and the Head of the Multimedia Research Group, and a member of the ‘Clepsydra’ Cultural Heritage Digitization Center. He has been an Adjunct Faculty Member of the Democritus University of Thrace, the University of the Aegean, and the Hellenic Open University. His research interests include 2D/3D imaging, CBIR, multimedia technologies, human–computer interaction, intelligent user interfaces, multi-sensory environments and ambient intelligence, 3D digitization and reconstruction, 3D-GIS, and mixed/augmented/virtual reality. He is a member of the Technical Chamber of Greece, of the Hellenic Researchers’ Association, and a Founding Member of the ‘Athena’ Research Center’s Researchers’ Association.

CHRISTOS KOULAMAS (Senior Member, IEEE) received the Diploma degree in computer engineering and informatics and the Ph.D. degree in electrical and computer engineering from the University of Patras, Greece, in 1994 and 2004, respectively. He has more than 25 years of experience in research and development, in the areas of real-time distributed embedded systems, industrial networks, wireless sensor networks, and their applications in various industry sectors. He is currently the Research Director of the Industrial Systems Institute (ISI), Athena Research and Innovation Center, Greece. He is the author or coauthor of over 70 scientific publications in international journals, conferences, and books, and of more than 110 publicly available or confidential technical reports. His current research interests include cyber-physical systems (CPS) and the Industrial Internet of Things (IIoT) technologies and their applications in smart environments.

Dr. Kalogeras has served as a Guest Editor for Electronics and Sensors journals (MDPI), as an Editorial Board Member of Sensors journal, as a PC/TPC member of many IEEE and other conferences and workshops, and as a member of the board of ERCIM (European Research Consortium for Informatics and Mathematics).

KONSTANTINOS MOUSTAKAS (Senior Member, IEEE) received the Diploma and Ph.D. degrees in electrical and computer engineering from the Aristotle University of Thessaloniki, Greece, in 2003 and 2007, respectively. From 2007 to 2011, he served as a Postdoctoral Research Fellow with the Information Technologies Institute, Centre for Research and Technology Hellas. He is currently an Associate Professor with the Electrical and Computer Engineering Department, University of Patras, the Head of the Visualization and Virtual Reality Group, the Director of the Wire Communications and Information Technology Laboratory, and the Director of the M.Sc. Program on Biomedical Engineering at the University of Patras. His main research interests include virtual, augmented and mixed reality, 3D geometry processing, haptics, virtual physiological human modeling, information visualization, physics-based simulations, computational geometry, computer vision, and stereoscopic image processing. In recent years, he has been the (co)author of more than 210 articles in refereed journals, edited books, and international conferences. His research work has received several awards. He is/was a Coordinator of the GameCar H2020 Project and a Scientific Coordinator of the NoTremor FP7 Project. He has also been a member of the organizing committee of several international conferences. He is a Senior Member of the IEEE Computer Society and a member of Eurographics. He serves as a regular reviewer for several technical journals and has participated in more than 20 research and development projects funded by the EC and the Greek Secretariat of Research and Technology.

ATHANASIOS KALOGERAS (Senior Member, IEEE) received the Diploma degree in electrical and engineering and the Ph.D. degree in electrical and computer engineering from the University of Patras, Greece, in 1991 and 2001, respectively. He has been with the Industrial Systems Institute, Athena Research and Innovation Center, since 2000, where he is currently the Research Director. He was employed as a Collaborating Researcher at the University of Patras, from 1991 to 1998, the Computer Technology Institute and Press DIOPHANTUS, from 2014 to 2015, and in the private sector. He was an Adjunct Faculty Member at the Technological Educational Foundation of Patras, from 2001 to 2015. He has over 25 years of research experience being involved in more than 45 research and development projects. His research interests include cyber physical systems and their sub class industrial control systems, the Industrial Internet of Things, industrial integration and interoperability, and collaborative manufacturing. His application areas include manufacturing environment, critical infrastructure protection, smart buildings, smart cities, smart energy, and tourism and culture.

Dr. Kalogoras has served as a program committee member for more than 30 conferences and as a reviewer for more than 40 international journals and conferences. He has been a member of the Board of Directors of AgriFood Western Greece, since 2019. He has been a Postgraduate Scholar of the Bodosakis Foundation. He is also a member of the Technical Chamber of Greece. He is a Local Editor of ERCIM News, Greece.