ARQGAN: an evaluation of Generative Adversarial Networks' approaches for automatic virtual restoration of Greek temples

Alberto Nogales[1], Emilio Delgado[2], Angel Melchor[1], and Álvaro J. García-Tejedor[1]

[1]CEIEC Research Institute, Universidad Francisco de Vitoria, Ctra. M-515 Pozuelo-Majadahonda km. 1,800, 28223 Pozuelo de Alarcón (Madrid), Spain
[2] Architecture School, Universidad Francisco de Vitoria, Ctra. M-515 Pozuelo-Majadahonda km. 1,800, 28223 Pozuelo de Alarcón (Madrid), Spain

alberto.nogales@ceiec.es, e.delgado.prof@ufv.es, angelmelchorsanchez@gmail.com, a.gtejedor@ceiec.es

Abstract. In the last years, Graphics Processing Units are evolving fast. This has had a big impact in several fields, such as Computer-Aided Design and particularly in 3D modeling, allowing the development of software for the creation of more detailed models. Nevertheless, building a 3D model is still a cumbersome and time-consuming task. Another field, that is evolving successfully due to this increase in computational capacity is Artificial Intelligence. These techniques are characterized among other things by the fact that they can automate tasks performed by humans. For example, reconstructing parts of images is being a hot topic recently. In this paper, a method based on Artificial Intelligence and in particular Deep Learning techniques is proposed to achieve this task. The aim is to automatically restore Greek temples based on renders of its ruins obtained from 3D model representations. Results show that adding segmented images to the training dataset gives better results. Also, restoration of the general part of the temples is well performed but the detailed elements have room for improvement.

Keywords: Neural networks; Deep Learning; Generative Adversarial Networks; Cultural heritage; Virtual restoration; Greek temples.

1 Introduction

European Commission surveys show that over 80% of the Europeans care about conserving their cultural heritage[1], with a particular interest in preserving historical sites as monuments or buildings. According to Berndt and Carlos (2000), the preservation of historical sites involves a range of different professionals, including art historians, archaeologists, restorers, or architects. It also includes a wide range of activities such as digitalization, preservation, or restoration. The latter is defined as "returning the existing fabric of a place to a known earlier state by removing accretions or by reassembling

[1] https://what-europe-does-for-me.eu/en/portal/2/B35

existing components without the introduction of new material", Jokilehto (2005). Restoration has undergone significant evolution in recent years due to technological advances like 3D models (either digital or physical), Virtual/Augmented Reality, and other tools. For example, 3D modeling has benefited from the drop in the cost of hardware components. Graphics Processing Units (GPUs) are increasingly affordable and offer enormous computing power. Nevertheless, the implementation of a 3D model is still complex and highly time-consuming. Besides, 3D models are specific for the building to be restored, there is no generic methodology or general-purpose tool. Each restoration process starts from scratch.

In parallel, the problem-solving capacity and applicability of Artificial Intelligence, specifically Deep Learning techniques, have improved. Deep Learning is defined in Lecun, Bengio & Hinton (2015) as models that can learn representations of data with multiple levels of abstraction. Deep Learning models are used in a wide variety of fields, including image processing, with good results, O'Mahony et al (2019). In this sense, the so-called image inpainting process can help in the restoration issues raised above.

The image-to-image translation is defined in Isola et al (2018) as "translating one possible representation of a scene into another, given sufficient training data" and comprises different techniques, such as style transfer, colorization, or image inpainting. The latter is defined in Yu et al (2018) as "synthesizing visually realistic and semantically plausible pixels for the missing regions that are coherent with existing one". Image inpainting is a good candidate to develop a method which, starting from an image of a monument or building in ruins, automatically obtaining an image of its possible restoration. This use of image inpainting techniques in this context can be called "virtual restoration".

This paper describes a method that automatically generates images of complete 3D models of buildings from images of partially destroyed 3D models. In particular, it covers the use case of virtually restoring Greek temples of different styles. This is a particular and innovative use case of inpainting as the reconstructed part has not previously be delimited like in previous works. The reconstruction is also particular as it has to infer details of the context as shadows or lightning. So, the model has to detect which part and what particular details need to be reconstructed. The method relies on the use of Generative Adversarial Network (GAN), a particular type of Deep Networks.

Two different training approaches have been followed: In the first one, the model has been trained with pairs of images of different ruin levels, from different visual perspectives, and the corresponding images of complete temples. The input in the second approach also uses an image of the complete temples with a color code for each different architectural element (a segmented image).

The work has been carried out by an interdisciplinary group composed of architects and computer science researchers. It comprises two stages: first, the creation of the dataset using a Computer-Aided Desing (CAD) program, and, second, the implementation and training of the GAN models. Finally, the results have been validated by using some mathematical metrics and surveys of students and professionals in the field. Results

show that segmented training with a combination of ruins and segmented images obtains results that are close to the original images. Also, it is concluded that training datasets must be improved for proper restoration of the details of the temples.
From a scientific perspective, the methodology set out in this project would make it possible to approach the reconstruction process in a different way to the procedures of 3D scanning, photogrammetry, and 3D modeling which, at the time of this study, are the only methods for tackling this issue. The definition and systematization of architectural languages and their integration in a neural model would allow users to face reconstruction projects in a more agile way, being able to obtain diverse assumptions modifying the codification of the architectural language and the theoretical foundation from the archaeological rest.
The rest of the paper is structured as follows. Section 2 gives a brief description of some related works. Section 3 explains the data and methods used for experimentation. Section 4 discusses the results obtained after the experimentation. Finally, section 5 summarizes the conclusions obtained through the work.

2 State of the art

Among various types of data, image restoration is perhaps the application that has been most widely studied. A method for image restoration is presented in Kumar et al (2019), using a Dual ascend based median filter in images with noise and blur. Also, Mairal et al (2019) implement a method that combines non-local means and sparse coding approaches to restoring images. In the case of actual paper, image restoration is solved by using Deep Learning techniques.
Some previous papers are using Deep Learning in image restoration that should be highlighted. For example, Chen et al (2019) present a framework for image restoration and recognition which uses CNNs with convolutional and deconvolutional layers. Another work that implements the restoration case of image deblurring is solved in Kupyn et al (2019). In this case, the model uses GANs architecture.
Finally, some works that use Deep Learning for image inpainting are listed given that this is the task to be solved in this paper. In Nezeri et al (2019), a two-step method is presented: first, it creates a sketch drawing the missing edges, then the rest of the image is filled in. Another similar work is Jo and Park (2019), here GANs are used to reconstruct images with missing parts that have been sketched by users. Another Deep Learning two-step method is presented in Wu et al (2019), in this case, portrait images are reconstructed: first inferring the human-body pose and, then, completing the image. Sun et al (2019) also present a two-stage method that comprises content reconstruction and texture detail restoration. Finally, Wang et al (2019) propose an extended image inpainting process applied to video where the model is a combination of 3D CNN with 2D CNN architectures.
As can be seen, many papers are dealing with this subject, this however is the first one using images of ruins of Greek temples and their complete versions as a training dataset. The proposed models have to identify the part of the image to be reconstructed as it has not been previously marked. It also should be taken into account that training has been

done in two different ways. Results from this work will let users speed up 3D modeling tasks that need specific knowledge and a great deal of time.

3 Materials and methods

3.1 Training dataset

This dataset was created manually by a student of the Degree of Architecture. The software used in this case is SketchUp[2] 2020, a 3D modeling program with applications in architecture, video game design, or mechanical engineering. This decision was based on the impossibility of having pairs of photographs of restored temples and their previous ruins kept in perspective. This is an important point in the fact of working only with synthetic images. Therefore, 3D digital models were rendered and images were obtained from them.

From an architectural perspective, some difficulties have been taken into account when designing the dataset. The first one is the existence of background (sky and landscape) behind the building. In general, any photograph of a building has a background that frames and contextualizes its presence. In the dataset, the relationship between the background and the building has cared so that the analysis of the model focuses on the building and not on the interpretation that could be produced by superimposing certain elements with surrounding objects. The second refers to the presence of scaffolding and auxiliary structures. Some of the reference buildings are in the process of being restored and the photographs analyzed present the provisional constructions, such as scaffolds, which are logically necessary to undertake these tasks. In the dataset, the presence of these elements has been omitted to avoid interference and focus the learning on the building and its parts. The third one has to do with the deteriorated areas and the stains of the different elements of the construction. In general, the architectural remains analyzed show an important material heterogeneity as a result of the passage of time, such as efflorescences, deterioration, due to partial destruction or the incidence of water, among others, and specific restorations that vary the material aspect of the stone. To facilitate the analysis of the images, in the dataset a homogeneous material has been used and respected the general integrity of the architectural elements. That is, when the assumptions of ruin have been designed for each of the case studies, complete architectural elements have been eliminated, such as an abacus, a section of the shaft, a triglyph, or a section of the entablature supported by two columns. The fourth refers to the particular ornamental elements of the building such as sculptures and reliefs. In the case of the Doric order, the metopes and triglyphs are sculptural elements that are part of the architectural language of the building. The latter are systematically repeated in the buildings that use this language and for this study, they have been taken into account. However, there are other elements, such as gargoyles, acroters, or the sculptural reliefs of pediments, which, due to their complexity, have not been taken into account in the dataset. The fifth difficulty arises when the images show interspersed or superimposed elements, such as people, vegetation, or fragments of the building that are not in their

[2] https://www.sketchup.com/

original position. For the construction of the dataset, these elements have been eliminated to visualize the building more clearly. The sixth refers to the encounter with the ground. In general, any building presents some singularity in the way it stands on the ground, making the stylobate and the stereobate manifest a special configuration. It has been assumed a horizontal terrain around the building. Finally, the seventh difficulty concerns the photograph itself, which is the object of analysis. Both the distortions produced by the camera or by the observer's point of view, as well as the color aberrations that may occur in the photograph, may make it difficult, in the first instance, for the neural model to analyze it. For this reason, the dataset has been constructed using perspectives with the points of view of a person, in which the building is perfectly framed within the images. Also, the use of a lighting system that generates a very contrasting spectrum of light and shadow has been avoided.
These seven aspects also justify the need to build a synthetic image dataset using autonomous 3D models that reproduce the analyzed buildings in the clearest way. To this end, a study has been carried out on the proportion and geometry of the different elements that make up the classical order, and the different cases have been constructed, taking existing Greek temples as a reference. It should be noted that it is quite difficult to standardize a style of architecture given the many nuances related to location, age, or civilization. Also, due to the interest in systematizing the architectural language, work has been done on a matrix of assumptions that makes the configuration of buildings plausible, taking into account the conceptual and geometric structure of said language. Fig. 1 shows the possible combinations of the Greek temples.

		Position of the columns							
		In Antis	Double Antis	Proprostyle	Amphipropter	Pseudoperipter	Peripter	Pseudodipter	Dipter
Classification	Distyle	X	X	-	-	-	-	-	-
	Tetrastyle	X	X	X	X	O	-	-	-
	Hexastyle	O	O	O	O	X	O	O	-
	Octastyle	-	-	-	-	X	X	X	X
	Aeneastyle	-	-	-	-	X	X	X	X
	Decastile	-	-	-	-	O	X	X	X
	Dodecastle	-	-	-	-	O	O	O	X

X: existing, probable; O: rare; -: impossible

Fig. 1. Matrix of the possible combinations of the columns' positions and styles for Greek temples.

To reinforce the learning of the neural network, the synthetic images are accompanied by others of a purely analytical nature using a palette of flat colors that are superimposed on each element that makes up the architectural language. These colors facilitate the identification of the architectonical elements of the building from different perspectives and points of view, making the reconstruction more accessible.
A total of 10 different temples have been modeled at 4 different states: the complete temple and three different states of conservation of the ruin. Then, each model was

rendered using V-Ray[3] NEXT version, a computer-generated rendering software. Rendering has been carried out from different perspectives, and the results are pairs of the complete temples with each of theirs remains in each perspective. This process was carried out by establishing eleven equally separated points around the building with a camera moving from one point to another obtaining 25 captures.
At the end of this process, the camera has gone around the whole temple obtaining 300 images, each one with a different perspective. As each temple can be found in 4 states, this makes a total of 1,200 images for each temple. Finally, as 10 different temples have been modeled, the dataset will have 12,000 different images. The size of the training set is related to the limitation of examples of Greek temples, as it has been remarked before. Although the training set seems small, there are other successful works with similar amounts of images, Elharrous et al (2019).
As the research follows two approaches called direct and segmented to train the models, there will be only two different training datasets. The first will be used for direct training: only pairs of images of the temple in ruins (input) and its restoration (output) are needed, both having the same perspective and lighting, obtained using Global Illumination and Sun&Sky System rendering options. The second dataset will be used for segmented training. In this case, another image was added making a triplet for each temple: the temple in ruins and the complete temple segmented by architectural elements (inputs) and the restored temple (output). Segmentation of the restored temple is obtained using the "material_id" channel to assign different colors to the different architectural elements of the temples. This segmentation encompasses well-known elements such as columns or friezes to more detailed features like triglyphs or metopes. As a segmented image must be obtained for the complete temple for each perspective, the segmented dataset will have 3,000 more images, corresponding to 300 views of each of the 10 different temple models.
An example of an instance for both training datasets can be seen in Fig. 2 and Fig. 3.

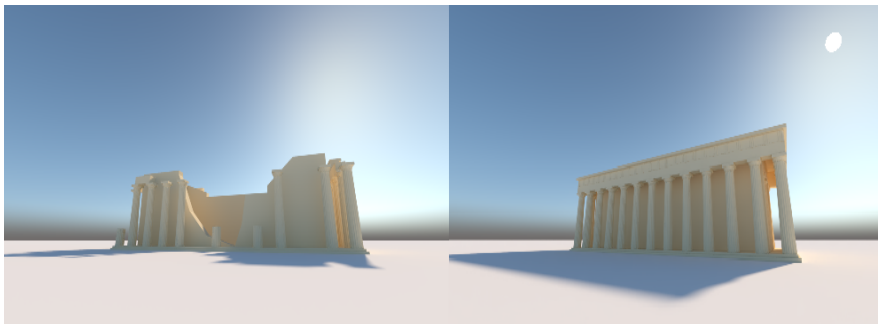

Fig. 2. A pair of images used in direct training.

[3] https://chaosgroup.com/vray/skecthup

For both direct and segmentation-aided training, the original datasets were split into three subsets: train, evaluation, and test. For the test set, 1,200 pairs of encompassing all the information about a particular temple were held out, i.e. images of the reconstructed version and its three possible ruin states. Thus, for training and evaluation, 10,800 pairs of images were used, split in 75%-25%. This division was done at the temple level, extracting a continuous set of perspectives of every model, different for every temple. A whole set of images of a particular temple was held out for testing purposes.

3.2 Data preprocessing

Firstly, the images require several transformations by coding a Python script using TensorFlow[4] which is an open-source platform for machine learning. First, they were downscaled to half their original size; then randomly cropped to fit the model's input shape (256x512) and subject to jittering in the form of random brightness, contrast, saturation, and horizontal mirroring. Before being fed into the model, the pixels' values were normalized from the original [0, 255] range to [-1, 1] to speed up the model's convergence.

3.3 Generative Adversarial Networks

The Deep Learning model proposed in this research is called GAN. They were first introduced in Goodfellow et al (2014) and consist of two neural models that compete between them. The first model, called generator tries to learn the data distribution of the training dataset in an unsupervised way to generate new instances. The second model is called discriminator and determines if the new data belongs to the training dataset or has been created by the generator model. Thus, the competition consists of the generator making the discriminator believe that an instance artificially generated (fake) belongs to the training set (real). This competition leads both models to improve their skills until the generated instances are practically indistinguishable from the original ones.

3.4 Inspirational architectures

As mentioned before, GANs comprise two neural architectures. Based on the results obtained with pix2pix, it has been decided to use a similar architecture, Isola et al (2017) for both the generator and the discriminator. The former uses a model based on a 2D Convolutional-Deconvolutional Autoencoder and the latter a Markovian discriminator.

Convolutional-Deconvoultional Autoencoder with 2 Dimensions. Autoencoders were first introduced by Hinton and Salakhutdinov (2006). The model reduces the dimensionality of the input data to an essential representation that, later, is upsampled obtaining the input data. This architecture was used in the generator using an input layer connected through several layers, encoder, to reduce the information and extract the

[4] https://www.tensorflow.org/

main features of the input data. Then, the feature representation is introduced in a decoder that upscales it to an output of the same dimension (generally) as the input data. This fits perfectly with the proposed training methods as the dataset is formed by pairs or triplets of images. In this case, the input data is the image of a ruined temple or this image plus the segmented one and the output is its reconstruction.
However, these architectures have a problem produced by the bottleneck, the part that connects the encoder with the decoder, just at the moment that the feature representations are obtained. It could occur that some features could not be transmitted. This is solved using Skip Connections. These were introduced by U-Net architectures in Karimov et al (2019) to establish connections between non-sequential layers from the encoder to the decoder. This will allow avoiding the problem of lost features caused by the bottleneck.
Another important contribution produced by U-Net is that of using convolutional layers in Autoencoders. Convolutional Neural Networks with 2 Dimensions was a big milestone in Deep Learning, Lecun et al (1999). Krizhevsky et al (2012), applied convolutional layers allowing the delocalized extraction of the main features of an image. In this way, a feature found in a part of an image can be found in another part in another image. In combination with the U-Net architecture, this will allow users to obtain the main features of the input images concentrated in a reduced data structure to be obtained and upsampled later.
Markovian Deconvolutional Networks. In contrast to many GAN models, the discriminator used in this case does not evaluate the generated image as a whole, but rather evaluates different patches separately. The benefit of using this discriminator is the evaluation of little patches as real or fakes, being able to analyze local textures. As a sliding window goes through the whole image, local continuity is taken into account in the analysis, and details of the context are easier to be found. The model creates an *NxN* patch that goes through the whole image giving an evaluation, fake or real, of each part of the image. This is based on the idea raised by Markovian Random Fields (MRFs) which assumes that the most relevant dependencies in an image can be found at the local level.
Summarizing, the discriminator will consist of a 2D CNN responsible for feature extraction. Then, an MRF patch is used to evaluate these features, obtaining a matrix with values from 0 to 1, where 0 means that this part of the image is synthetic and 1 belonging to the original dataset. The final evaluation of the image determining if it is real or fake is described at the training phase.

3.5 The proposed solution

The problem raised in this paper comprises two different approaches, both implemented by developing a Python script with TensorFlow[5]. The first approach uses a straightforward training and can reconstruct images of Greek temples directly from its ruins. The second approach does the same reconstruction aided by additional information in the form of a segmented image. These two different solutions need three different architectures, all based on GAN models.

[5] https://www.tensorflow.org/

Generator for temple restoration. The two training approaches are very similar. This implies two different architectures depending on the input data although both work in a similar way. An image of the ruins is fed to the model for direct training. In the case of using segmentation, the ruined temple is used alongside with the segmented image of the temple. The size of the input data is reduced by applying convolutional filters until the main features are extracted. Then, these feature maps are upscaled and converted back into an image, a complete temple. This image is then fed to the discriminator that evaluates it; with that comparison the generator learns how to improve its results, tuning both architectures its parameters.

The Input Layer is the only difference between the models used in the direct training and segmented training. The former has a single Input Layer and the latter uses both the image of the temple ruins and the image of its segmentation as input data, this entails modifying the inputs of the model. For the first approach, the Input Layer has the size of a 512x256x3 colored image. For the second one, two images with the previous size are concatenated in an input of 512x256x6.

From this point on, both architectures remain the same. The Input Layer is then connected to a set of Convolutional Blocks whose aim is to reduce the dimensionality of the data. Each block, except for the first one, is composed of a 2D Convolutional Layer, a Batch Normalization layer, and Leaky ReLU as the activation function. The number of neurons varies in some of the blocks. It is increased in the first four, starting with 64 and then going to 128, 256, and 512; therefore other four blocks that remain with 512 neurons each.

After reducing the image, the Deconvolutional stage begins. It aims to up-sample the feature representation of the images to an output data with the same size as a single image. This stage is composed of seven Deconvolutional Blocks, one less than in the previous stage. Each of the blocks is composed of a Transposed 2D Convolutional layer, a Batch Normalization layer, a Dropout layer, and Leaky ReLU as the activation function. The number of neurons is 512 for the first four blocks and then decreasing to 256, 128, and 64 each block. During this stage, an output image is created by up-sampling the essential features extracted from the input data. After that, a final 2D Convolutional layer with 3 neurons is applied to map the output data to a structure with RGB channels. This layer uses hyperbolic tangent as an activation function, so values are in the range of 0 to 1.

But these Convolutional-Deconvolutional models can cause problems since important data is lost when reducing input data to extract features. This was solved by U-Net models by implementing what is called Skip Connections. Thanks to these connections, some features that can be lost during the reduction process can be fed back in. Fig. 1 and Fig. 2 in Appendix A represent the architectures used in the generators of direct and segmented training.

Temple image discriminator. The second part corresponds to the discriminator, which in this case will have the same architecture for both types of training. First of all, there are two Input Layer with dimensions of 256x512x3 one corresponding to the generated image and another to the real image. Then, there is a Concatenation Layer that stacks both inputs generating and output data of 256x512x6. At this point, a set of Convolutional Blocks is used. Each block has a 2D Convolutional Layer, a Batch Normalization

with Leaky ReLU as the activation function. In total there are four blocks with an increasing number of neurons. The first block contains 64 and then 128, 256, and 512. After that convolutional process, there is a 2D Convolutional Layer that uses a sigmoid activation function and a filter of dimension 4x4 with stride 1. The output of the model is a 29x13 matrix whose values range from 0 to 1. Each of these values determines the validity of a patch from the input image; depending on if the patch of the image has been considered fake or real. Fig. 3 of Appendix A describes this architecture.

3.6 Training phase

Finally, the models have to be trained to obtain a set of hyperparameters that perform well. Even though both approaches are trained separately, they only differ in the used input data.

The training process is repeated iteratively in a way that generator and discriminator compete between them, tuning their hyperparameters at the same time. In terms of time: the first type of training needs 4 hours and 7 minutes and the second needs 11 hours and 37 minutes. The training is based on the idea of obtaining a probability that evaluates the synthetic image as a whole by applying binary cross-entropy to an output matrix.

Let $x \in X$ be an input where x is the image of a temple in ruins for the direct training or an image of ruins with its segmented image in the segmented training and X the training dataset. Let $y \in Y$ the expected output where y is a complete temple and Y the dataset of complete temples. G reconstructs an image x_{gen} which can be denoted as $G(x)=x_{gen}$. Then, the real image and the generated image are fed into the discriminator giving an output value which is a matrix of 0's and 1's, $D([x,y])=m_{real}$ and $D([x,x_{gen}])=m_{gen}$. At this point, the error produced by G and D has to be calculated. The loss function produced by G is given by applying Equation 5.

$$G_{loss}=BCE(1_{MxN},m_{gen})+\lambda*mean(\|y-x_{gen}\|_1) \quad (5)$$

In the previous equation *BCE*, which stands by Binary Cross-Entropy, is calculated as in Equation 6. *BCE* is applied to the output matrix from the discriminator obtaining values in the range from 0 to 1. The closer the value is to 1, the more real the image created will appear. In this Equation, y is the expected value and x the predicted.

$$BCE(y,x)=max(x,0)-x*y+log(1+e^{-|x|}) \quad (6)$$

In the case of the loss of the generator, *BCE* is applied using m_{gen} and a matrix of 1's with the same dimension as m_{gen}. Then, λ is used alongside the L_1 distance between the expected image and the generated one. The usage of λ, whose value is 100, in this context reduces generated noise artifacts, Isola et al (2017).

During the training process, the generator's loss has to be used with the loss produced by the discriminator. This is calculated by applying Equation 7.

$$D_{loss}=0.5*BCE(0_{MxN},d_{gen})+BCE(1_{MxN},d_{real}) \quad (7)$$

In this case, the *BCE* is calculated using the output of the discriminator and a matrix of zeros with the same dimension plus the output of the discriminator after introducing the generated image with the real image and a matrix of 1's with the same dimension. Once both losses are obtained, the weights of the neural models are updated using an optimizer. In this case, Adam, "an algorithm for first-order gradient-based optimization of stochastic objective functions, based on adaptive estimates of lower-order moments", Kingma and Ba (2014).
Both trainings approaches were performed in an Intel Core i5-8500 CPU@ 3.00GHz (6 CPUs), 32768 RAM MB with NVIDIA GeForce RTX 2080 Ti which lasted 4 hours and 7 minutes for direct and 11 hours and 37 minutes for the segmented one.

4 Results and evaluation

4.1 Objective evaluation

These evaluations are made using mathematical metrics, software or other methods that do not involve people to prevent personal biases. The approach applied is called Classifier Two-Sample Test (C2ST) , found in Lehmann and Romano (2006). The main aim of this test is to determine if two samples belong to the same distribution or, to check if two distributions, P and Q, are equal. In consequence, the null hypothesis stands for $P=Q$ and the alternative hypothesis for $P\neq Q$.
This can be achieved by evaluating the generator by taking into account the performance of a new discriminator. This consists of creating a new test set divided into two subsets: test-training and test-validation. The first subset is used to train the new discriminator whose aim is to discriminate between real and fake images. This discriminator is then tested with the test-validation subset with the target metric. Intuitively, if $P=Q$, the accuracy should be approximately 50% for a binary classifier. If by contrast $P\neq Q$, accuracy will be far from the previous value. After conducting an evaluation with 5-fold cross-validation (making the evaluation by splitting the test-validation dataset into 5 subsets). Table 1 shows the results for both experiments with a satisfactory result of near 54% accuracy for segmented training, concluding its better performance comparing to the direct training.

Table 1. K-fold cross-validation in both trainings

K	Direct Training	Segmented Training
1	69.67%	54.56%
2	57.50%	51.83%
3	55.54%	52.47%
4	64.16%	51.71%
5	60.14%	62.76%
Average	**61.40%**	**54.66%**

4.2 Subjective evaluation

As the results of this work are closely related to human perception, the performance of the model has been measured using personal opinions. This is considered highly subjective given that it depends on factors such as field of specialisation, experience, etc. The results are statistics applied to the data collected by a human survey of 18 people. The survey was made online by using Google Forms and it consisted of 4 parts. First, a set of 15 single images were shown. These images correspond to three possible cases: 5 were images that correspond to the training dataset (images of complete temples), 5 were generated images that correspond to the direct training and 5 were images created with the segmented training. During the second stage, 10 pairs of temples in ruins and its restoration were shown, corresponding to the best training evaluation in the previous step. The third stage was an evaluation of the presented images as a whole. Finally, individuals could offer a free comment about their perception of the images.
The questionnaire consisted of 6 questions for the first and second stages, two for the third and a final comment for stage four. Each question had a single answer, scored from 1 to 5. In Appendix B, Tables 1 and 2 show the possible responses of the test participants. The experiment was conducted on desktop computers or laptops with unlimited time to look at the images and give an answer.
The results obtained are presented in three different graphs. The first evaluates the differences between images of complete temples generated by direct training, segmented training and those created manually. Fig. 4 provides the results, divided by columns depending on the answers to the questionnaire. Each column has three boxplots that compile the evaluation from 1 to 5 given by the 18 participants depending on the type of image.

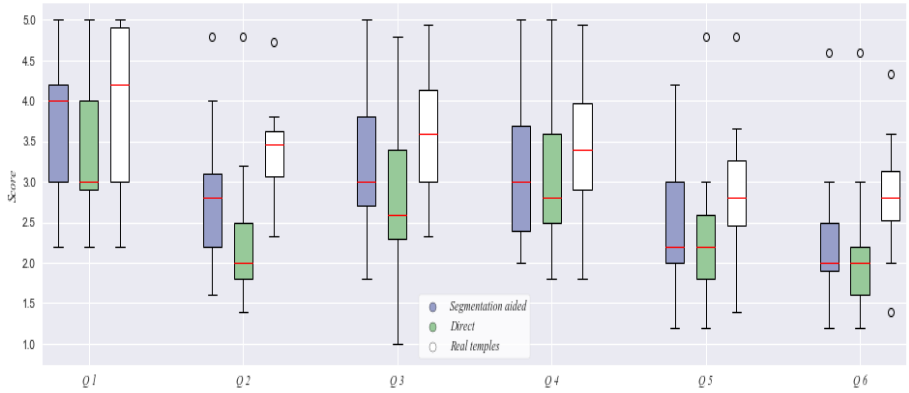

Fig. 4. Evaluation of images of complete temples. Each boxplot represents groups of images corresponding to the three different sets of images and the scores given in questions 1 to 6 by surveyed people.

As can be seen in the Fig. 4 above, the real images (white color) scored highest. It also should be noted that the segmented training (blue color) seems to perform better than the direct (green color). Another interesting point is that the worst evaluations for generated images (Q5 and Q6), also have poor values in the case of real images. These two

questions are related to the definition, quality, and resolution of the images, leading to the conclusion that this is an area for improvement in the training dataset. Question 2 also scores poorly for fake images but slightly better for real ones. In this case, the difference seems to be related to the details in the temples; again it can be concluded that the training dataset should be improved. Regarding the rest of the questions, the performance is almost the same with the exception of question 1. This shows the highest scores for real images and generated images with segmentation. As this question evaluates if the image seems real, it can be concluded that in general segmented restoration is performing well. As segmented training shows good results, the second part of the evaluation consisted of showing pairs of ruined and restored temples using this method. Results have been compiled in Fig. 5 shows a boxplot for each question.

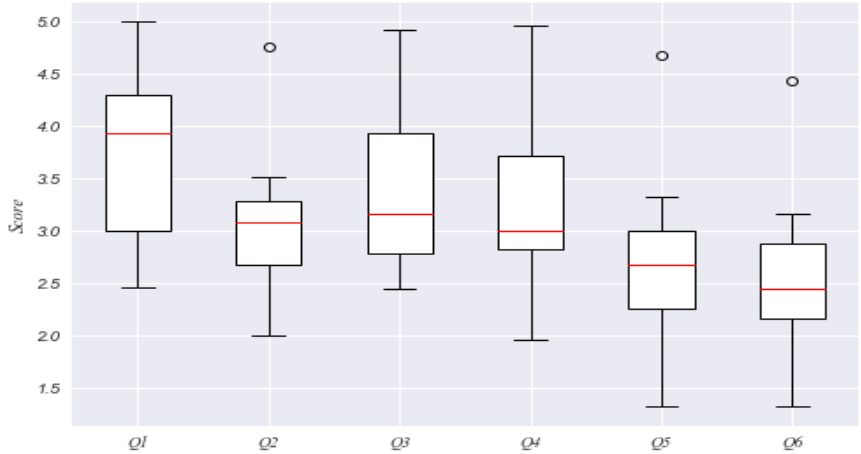

Fig. 5. Evaluation of pairs of images (ruined temples vs restored temples). Each boxplot collects the scores given in questions 1 to 6 by the surveyed people.

An analysis of Fig. 5 shows several things. At first sight and comparing with Fig. 4, evaluations are very similar: question 1 obtains the highest scores and questions 5 and 6 the lowest. Questions 2, 3 and 4 are in the mean with a value of 3 points out of 5. It can be concluded that restorations are performing well (Q1), the training dataset lacks the definition of details and image resolution (Q5 and Q6), and basic elements and architectonical order is well defined after restoration (Q2, Q3, and Q4). Finally, the evaluation was divided between biased (those having an explanation of the research) and unbiased individuals shown in Fig. 6.

Fig. 6. Differences between evaluations regarding biased and unbiased evaluators. Each pair of bars represents the scores given in questions 1 to 12 by the biased and unbiased groups of surveyed people.

Fig. 6 shows the evaluation regarding the questions for part 1 (questions 1 to 6) and 2 (questions 7 to 12) of the questionnaire. Notably, biased people make better evaluations of images of complete temples in isolation. In the case of pairs of ruined temples and their restoration, unbiased people gave higher scores than biased ones.
In the light of these results, it can be concluded, on the one hand, that people who knew the context of the project have not met the expectations set for restoration through Deep Learning models (Q7-Q12). On the other hand, it is also possible that biased people have been more critical in evaluating the responses in this last section. This problem is reinforced by the general opinion that this method cannot be extended to the analysis of other architectural languages. The impartial evaluators, without knowing the project, have maintained a homogeneous vision of the whole panorama. It could happen that they probably did not distinguish very clearly what was restored using a 3D model or a Neural Network.
Regarding the third part of the questionnaire, eleven of the surveyed people think (they evaluate this question with 3 or above) that this method could replace the actual methods for virtual restoration. The other question was regarding the application of this method to other architectural styles and eleven people gave three points or more.
The final part of the evaluation consisted of free comments. The recommendations received are related to aesthetic-compositional aspects of the images and technical aspects.

4.3 Qualitave evaluation

From an architectural point of view, the results are very positive. Certainly, the digital reconstruction is capable of representing the building envelope as it could do with other types of objects. However, the scope of this reconstruction also integrates particular aspects of the building that have to do with the architectural language used for its de-

sign. Analyzing the results obtained, it can be seen how the reconstruction of the fundamental parts of the classical order is coherent. Both the columns and the architrave are reconstructed in a differentiated and clear way. In addition, the parts that make up the previous elements can be identified. In the case of the columns, the reconstruction of the edges of the shaft, the subtle decrease of the section of the column as it rises from the podium and the presence of the capital with its fundamental parts, specifically the abacus and the equinus (the collar is not perceived due to the resolution of the images). The reconstruction of the entablature is less clear. Although the general shape of the architectural element (including the cornice) is respected, the line of the impost and the metopes and triglyphs are blurred in the resulting images. In general, the neural network is capable of shaping all the main elements that make the Doric style identifiable.

5 Conclusions and future works

The main aim of this work consisted of creating a Deep Learning model based on GANs for automatic digital reconstruction of Greek temples. In particular, GANs were used and trained with two different approaches: direct and segmented. The first problem was to generate manually two training datasets. To do that, 3D models in different states were obtained and images were captioned from then. Then, a set of instances for each dataset was created. In the first case, it consisted of pairs of ruined temples and its complete case. In the second dataset, an additional image was added of the complete temple segmented by colours. Subsequently, a GAN model formed by Convolutional/Deconvolutional Autoencoders (they differ in the Input Layer) for the generators and a Markovian Deconvolutional Network for the discriminator were created. Then, these models were trained in the two proposed ways: direct and segmented. At the end of the training, a set of temples was restored in the test stage. Finally, the images of restored temples were evaluated objectively and subjectively. Results in the first evaluation showed that the models perform well. In the case of the subjective test, it can be said that restoration was generally done well, but training datasets should be improved in details and resolution for better restoration of small elements.

In response to initial questions about the possibility that this work may serve as an alternative methodology to virtual restoration, the results of the experiments conducted are very positive. The fact of having trained the model with a series of assumptions and obtaining good results from images restored from scenes that have not been shown during the training makes thinking that this methodology can be very operative for the analysis of any type of temple of the Doria and Ionian stages. The rapid response of the model (milliseconds) can compete with traditional systems of three-dimensional restoration.

As future works, various improvements can be made. Datasets can be modified and introduced in the models in different orders, details in temples or image resolution can be improved. It will also be interesting to pay attention to the recommendations made by the people who have participated in the evaluation survey, described above, with regard to aesthetic, compositional and technical aspects. As segmented trained offers better results, a three-stage model should be developed, consisting of a set of GANs

that will receive a ruined temple that will be segmented, then a restored complete temple with coloured elements will be obtained, and finally, this will be transformed into an image similar to the 3D models. The models can be retrained using text guide by introducing descriptions of the Greek temples to be restored. Pictures of places can also be added, thus situating the temples in the real location where restoration is being carried out. Finally, a style transfer from 3D models to real pictures of temples could be achieved in order to give the temples a more realistic view.

Acknowledgments

This work is part of the ARQGAN project and was founded by the 2020 Call for Research Projects of the Universidad Francisco de Vitoria. The authors wish to thank Ignacio Barrera Muñiz for his help in preparing the Temples dataset.

References

1. Assael, Y., Sommerschield, T., & Prag, J. (2019, November). Restoring ancient text using deep learning: a case study on Greek epigraphy. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 6369-6376).
2. Berndt, E., & Carlos, J. (2000). Cultural heritage in the mature era of computer graphics. *IEEE Computer Graphics and Applications*, *20*(1), 36-37.
3. Chen, R., Mihaylova, L., Zhu, H., & Bouaynaya, N. C. (2019). A Deep Learning Framework for Joint Image Restoration and Recognition. Circuits, Systems, and Signal Processing, 1-20.
4. Elharrouss, O., Almaadeed, N., Al-Maadeed, S., & Akbari, Y. (2019). Image inpainting: A review. *Neural Processing Letters*, 1-22.
5. Esquef, P., Välimäki, V., & Karjalainen, M. (2001). Audio restoration using sound source modeling. In Proc. 2001 Finnish Signal Processing Symp. (FINSIG'01) (p. 47).
6. Ginsburgh, V. A., & Throsby, D. (Eds.). (2006). *Handbook of the Economics of Art and Culture* (Vol. 1). Elsevier.
7. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. In Advances in neural information processing systems (pp. 2672-2680).
8. Guner, A. F., & Benli, G. (2019). Project Management in Conservation and Restoration of Historic Buildings.
9. Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. science, 313(5786), 504-507.
10. Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167.
11. Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1125-1134).
12. Jo, Y., & Park, J. (2019). SC-FEGAN: Face Editing Generative Adversarial Network with User's Sketch and Color. arXiv preprint arXiv:1902.06838.

13. Jokilehto, J. (2005). Definition of cultural heritage: References to documents in history. *ICCROM Working Group 'Heritage and Society*, 4-8.
14. Karimov, A., Razumov, A., Manbatchurina, R., Simonova, K., Donets, I., Vlasova, A., ... & Ushenin, K. (2019). Comparison of UNet, ENet, and BoxENet for Segmentation of Mast Cells in Scans of Histological Slices. arXiv preprint arXiv:1909.06840.
15. Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
16. Kojoma, Y., & Washizawa, Y. (2018, November). Restoration of dry electrode EEG using deep convolutional neural network. In 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) (pp. 834-837). IEEE.
17. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).
18. Kumar, N., Shukla, H. S., Tiwari, A. K., & Dahiya, A. K. (2019). Dual Ascent Based Median Filter for Image Restoration. Available at SSRN 3350334.
19. Kupyn, O., Martyniuk, T., Wu, J., & Wang, Z. (2019). DeblurGAN-v2: Deblurring (Orders-of-Magnitude) Faster and Better. In Proceedings of the IEEE International Conference on Computer Vision (pp. 8878-8887).
20. Lehmann, E. L., & Romano, J. P. (2006). Testing statistical hypotheses. Springer Science & Business Media.
21. Lu, X., Matsuda, S., Hori, C., & Kashioka, H. (2012). Speech restoration based on deep learning autoencoder with layer-wised pretraining. In Thirteenth Annual Conference of the International Speech Communication Association.
22. Mairal, J., Bach, F. R., Ponce, J., Sapiro, G., & Zisserman, A. (2009, September). Non-local sparse models for image restoration. In ICCV (Vol. 29, pp. 54-62).
23. Mathieu, M., Couprie, C., & LeCun, Y. (2015). Deep multi-scale video prediction beyond mean square error. arXiv preprint arXiv:1511.05440.
24. Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784.
25. Nazeri, K., Ng, E., Joseph, T., Qureshi, F., & Ebrahimi, M. (2019). Edgeconnect: Generative image inpainting with adversarial edge learning. arXiv preprint arXiv:1901.00212.
26. LeCun, Y., Haffner, P., Bottou, L., & Bengio, Y. (1999). Object recognition with gradient-based learning. In *Shape, contour and grouping in computer vision* (pp. 319-345). Springer, Berlin, Heidelberg.
27. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444.
28. O'Mahony, N., Campbell, S., Carvalho, A., Harapanahalli, S., Hernandez, G. V., Krpalkova, L., & Walsh, J. (2019, April). Deep learning vs. traditional computer vision. In *Science and Information Conference* (pp. 128-144). Springer, Cham.
29. Pham, L. N., Tran, V. H., & Nguyen, V. V. (2013, November). Vietnamese text accent restoration with statistical machine translation. In Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27) (pp. 423-429).
30. Sijbers, J., Michiels, I., Van Audekerke, J., Verhoye, M., Van der Linden, A. M., & Van Dyck, D. (2000, June). Automatic EEG signal restoration during simultaneous EEG/MR acquisitions. In Medical Imaging 2000: Image Processing (Vol. 3979, pp. 1482-1491). International Society for Optics and Photonics.
31. Sullivan, A. M. (2015). Cultural Heritage & New Media: A Future for the Past. *J. Marshall Rev. Intell. Prop. L.*, *15*, 604.
32. Sun, T., Fang, W., Chen, W., Yao, Y., Bi, F., & Wu, B. (2019). High-Resolution Image Inpainting Based on Multi-Scale Neural Network. Electronics, 8(11), 1370.

33. Vecco, M. (2010). A definition of cultural heritage: From the tangible to the intangible. *Journal of Cultural Heritage*, *11*(3), 321-324.
34. Wang, C., Huang, H., Han, X., & Wang, J. (2019, July). Video inpainting by jointly learning temporal structure and spatial details. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, pp. 5232-5239).
35. Wu, X., Li, R. L., Zhang, F. L., Liu, J. C., Wang, J., Shamir, A., & Hu, S. M. (2019). Deep portrait image completion and extrapolation. IEEE Transactions on Image Processing.
36. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., & Huang, T. S. (2018). Generative image inpainting with contextual attention. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5505-5514).

Appendix A: Model architectures.

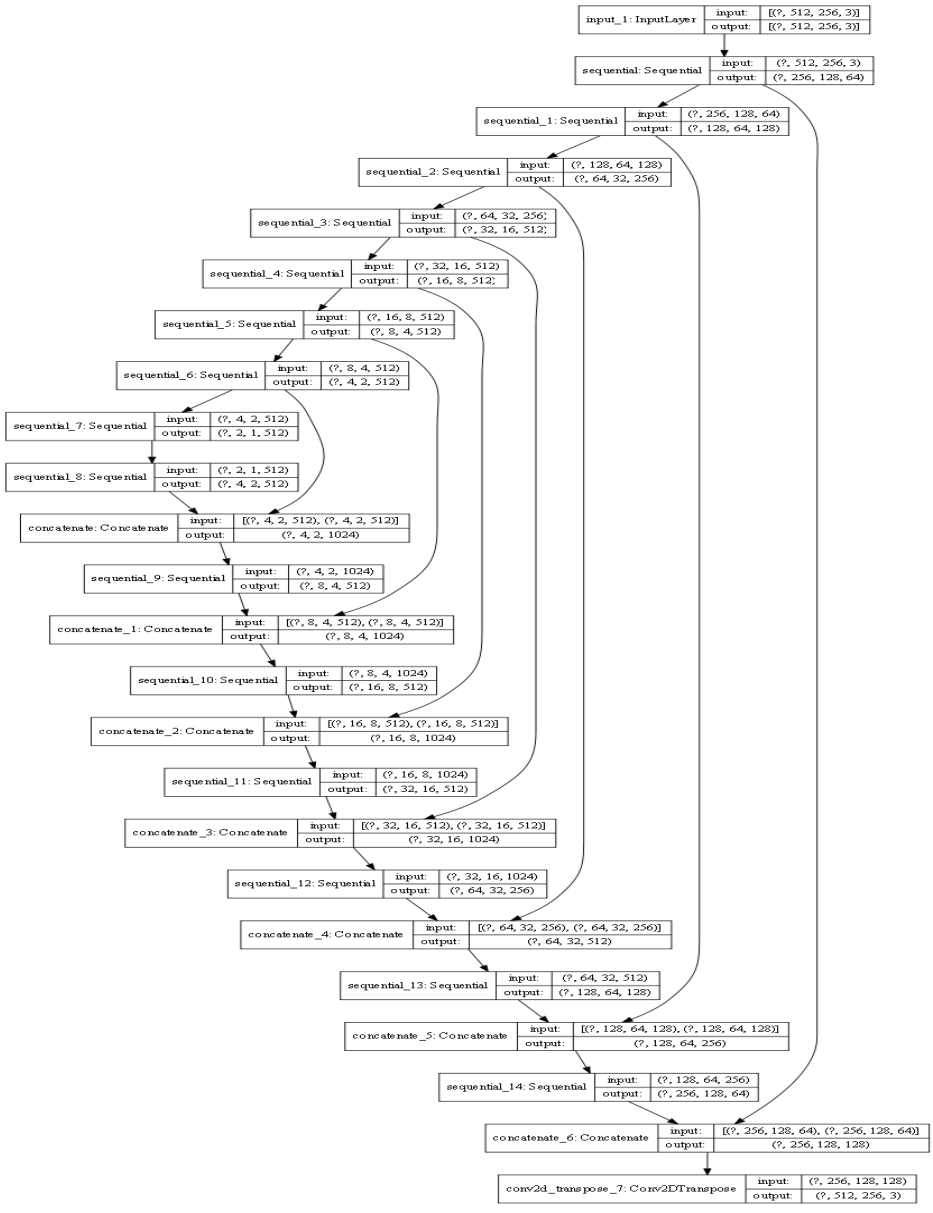

Fig 1. Generator's architecture for direct training.

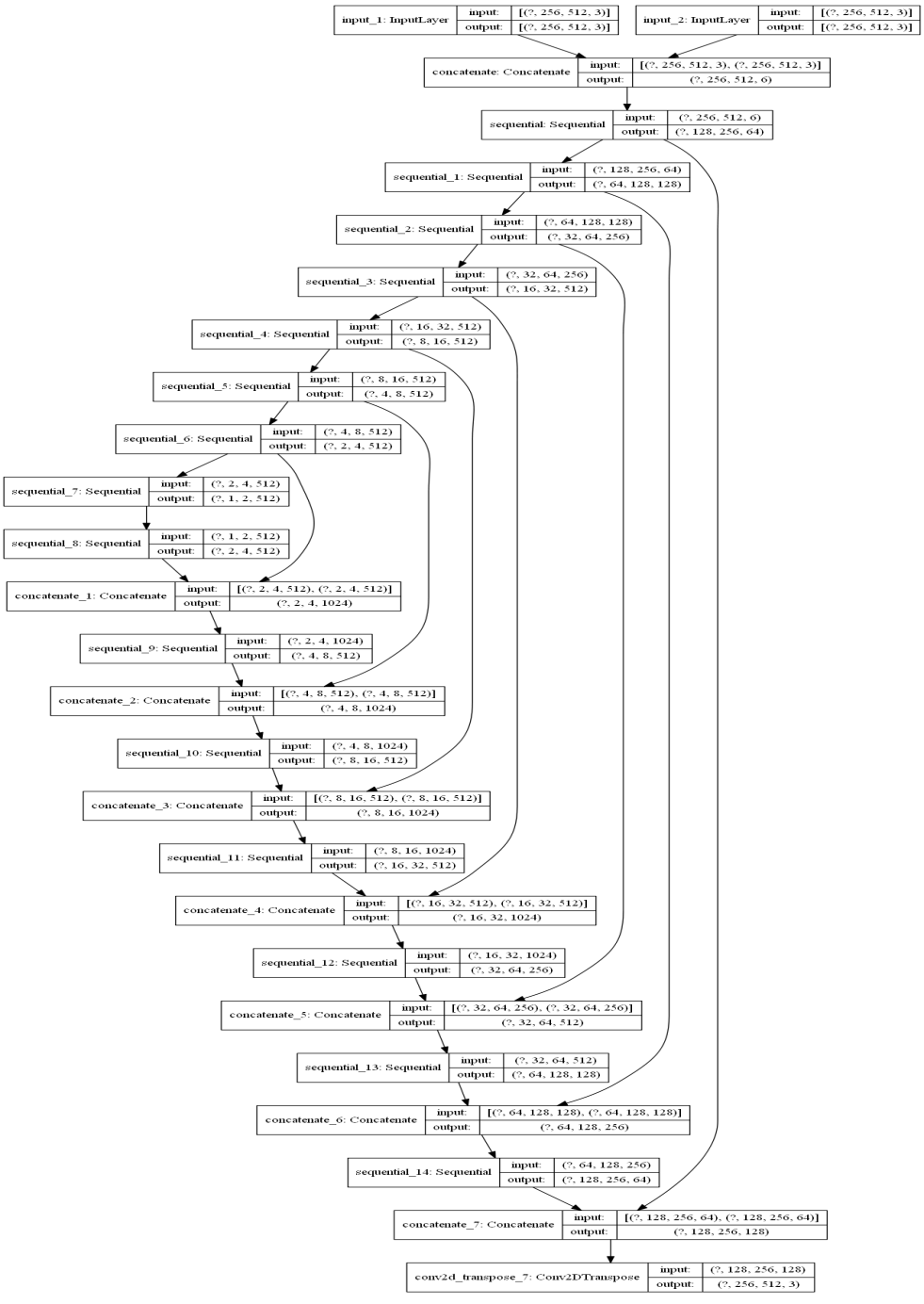

Fig 2. Generator's architecture for segmentation training.

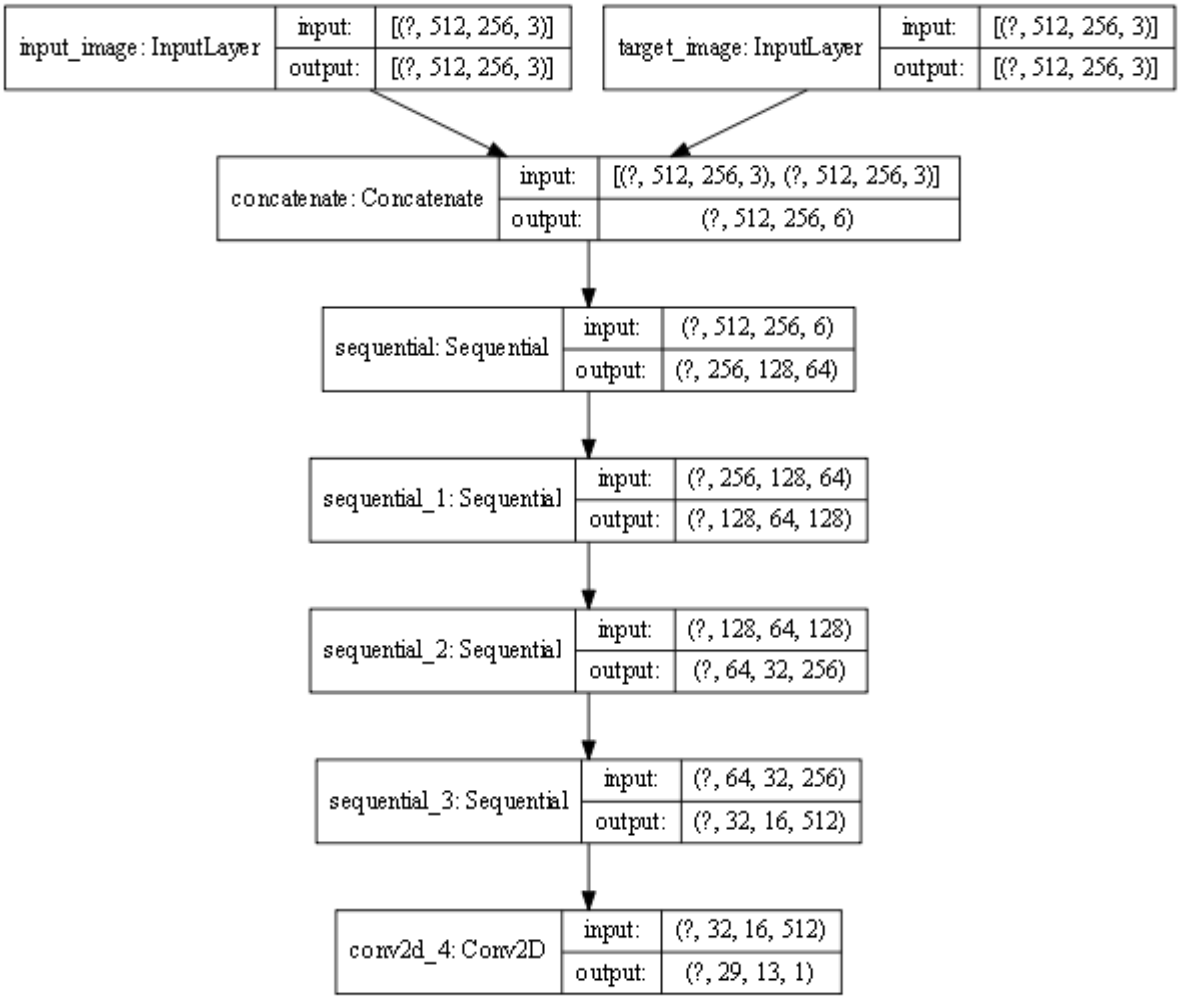

Fig 3. Discrimator's architecture.

Appendix B: Formulation of the questionnaires.

Table 1. Evaluation of images of complete/restored temples.

Evaluation items		1	2	3	4	5
1. Formal aspects	**1.1. First sight**	You don't understand what's going on.	You can guess a building.	You can clearly see a building.	Some parts of the image look very real.	The image presents a building that looks real.
	Can you see a building in the picture?	No				Seems real
	1.2. Basic construction	The results are not understood and/or there are significant inconsistencies in the formal setting.	An architectural construction can be intuited, but some important inconsistencies can be observed.	The building is evident and is completed with basic architectural elements.	The building includes constructive and ornamental details.	The building includes small nuances that complete the constructive and ornamental details.
	Are there any inconsistencies?	Yes, many				No, it's very good.
2. Identification of the architectural language	**2.1. Architectural coherence**	The result is not consistent with any known architectural language.	One intuits an architectural language, but on closer inspection, one sees that it does not correspond to known examples.	A classic architectural language can be sensed.	A classic architectural language is observed.	You can clearly see a classic architectural language.
	Is a coherent architectural language identified?	No				Yes

	2.2. Classical language	The architectural language in the reconstruction is not precisely understood.	The order (Doric or Ionian) is intuitive, but on closer inspection, it is clear that it does not correspond to known examples.	There is an order (Doric or Ionian) in the building.	You can clearly see an order (Doric or Ionian) in the building.	The building presents an order (Doric or Ionic) recognizable in the details and nuances.
	Can you identify a particular classical language?	No				Yes
3. Technical assessment	**3.1. Defining the details**	In the picture, the details are rough and confusing.	Although some detail is observed, the result incorporates too much noise when viewed in detail.	The details observed are reasonable in relation to the size of the image.	In general, all the main architectural elements of the building are visible.	The quality of the details allows for distinguishing small architectural elements.
	Do the details look good?	No				Yes
	3.2. Quality and resolution	The resolution and noise of the image make the result incomprehensible.	Being aware of the size of the resulting image, the quality of the image is sufficient to understand the intention, but insufficient to identify more fine aspects.	The quality and resolution of the result are acceptable.	The resulting image is noise-free and the reconstruction is clearly visible.	The nuances and ornamental details are clearly shown.
	How do you evaluate the resolution?	A lot of noise				It looks clear

Table 2. Evaluation of images of temples in ruins and its restoration.

Evaluation items		1	2	3	4	5
1. Formal aspects	**1.1. First sight**	You don't understand what's going on.	Two images can be intuitively related in a diffuse way.	An image of a previous state and an image of a reconstructed state are clearly perceived.	In the first impression, the reconstruction is evident and, from a formal (architectural) perspective, it is coherent.	The reconstruction looks real.
	Looking at the pair of images, do you understand a reconstruction?	No				Seems real
	1.2. Basic construction	The results are not understood and/or there are significant inconsistencies in the formal setting.	An architectural reconstruction can be intuited, but some important inconsistencies can be observed.	The reconstruction is evident, completing the basic architectural elements.	The scope of the reconstruction includes constructive and ornamental details.	The reconstruction includes small nuances that complete the constructive and ornamental details.
	Are there any inconsistencies?	Yes, many				No, it's very good.
2. Identification of the architectural language	**2.1. Architectural coherence**	The result is not consistent with any known architectural language.	One intuits an architectural language, but on closer inspection, one sees that it does not correspond to known examples.	A classic architectural language can be sensed.	A classic architectural language is observed.	You can clearly see a classic architectural language.
	Is a coherent architectural language identified?	No				Yes
	2.2. Classical language	The architectural language in the reconstruction is not precisely understood.	The order (Doric or Ionian) is intuitive, but on closer inspection, it is clear that it does not correspond to known examples.	There is an order (Doric or Ionian) in the building.	You can clearly see an order (Doric or Ionian) in the building.	The building presents an order (Doric or Ionic) recognizable in the details and nuances.
	Can you identify a particular classical language?	No				Yes

3. Technical assessment	**3.1. Defining the details**	In the picture, the details are rough and confusing.	Although some detail is observed, the result incorporates too much noise when viewed in detail.	The details observed are reasonable in relation to the size of the image.	In general, all the main architectural elements of the building are visible.	The quality of the details allows for distinguishing small architectural elements.
	Do the details look good?	No				Yes
	3.2. Quality and resolution	The resolution and noise of the image make the result incomprehensible.	Being aware of the size of the resulting image, the quality of the image is sufficient to understand the intention, but insufficient to identify more fine aspects.	The quality and resolution of the result are acceptable.	The resulting image is noise-free and the reconstruction is clearly visible.	The nuances and ornamental details are clearly shown.
	How do you evaluate the resolution?	A lot of noise				It looks clear