Retro-Remote Sensing: Generating Images From Ancient Texts

Mesay Belete Bejiga, Farid Melgani, *Fellow, IEEE*, and Antonio Vascotto

***Abstract*—The data available in the world come in various modalities, such as audio, text, image, and video. Each data modality has different statistical properties. Understanding each modality, individually, and the relationship between the modalities is vital for a better understanding of the environment surrounding us. Multimodal learning models allow us to process and extract useful information from multimodal sources. For instance, image captioning and text-to-image synthesis are examples of multimodal learning, which require mapping between texts and images. In this paper, we introduce a research area that has never been explored by the remote sensing community, namely the synthesis of remote sensing images from text descriptions. More specifically, in this paper, we focus on exploiting ancient text descriptions of geographical areas, inherited from previous civilizations, to generate equivalent remote sensing images. From a methodological perspective, we propose to rely on generative adversarial networks (GANs) to convert the text descriptions into equivalent pixel values. GANs are a recently proposed class of generative models that formulate learning the distribution of a given dataset as an adversarial competition between two networks. The learned distribution is represented using the weights of a deep neural network and can be used to generate more samples. To fulfill the purpose of this paper, we collected satellite images and ancient texts to train the network. We present the interesting results obtained and propose various future research paths that we believe are important to further develop this new research area.**

***Index Terms*—Convolutional neural networks (CNN), deep learning, generative adversarial networks (GAN), multimodal learning, remote sensing, text-to-image synthesis.**

I. Introduction

THE development of remote sensing technologies has enabled us to acquire information about objects on the Earth's surface from a distance, without a direct contact with the object under study. Analyzing this information helps us to better understand our environment. Developing efficient and effective methods to uncover meaningful information from these data is a long-standing goal for the remote sensing community. Over the years, the community has proposed several methods for the analysis of satellite and aerial remote sensing data and used them for a wide range of applications, such as vegetation monitoring, urban mapping, land cover/land use classification, and change detection.

The history of remote sensing dates back to 1859 when Gaspard Tournachon used balloons to acquire images of a small village near Paris [1]. During the civil war in 1862, the United States army used aerial photography for reconnaissance missions. The invention of airplanes in 1903 brought an alternative acquisition platform to balloons to aerial photography for reconnaissance missions during the First World War. Between World Wars I and II, civilian applications such as cartography, agriculture, and forestry started to use aerial photography. The development of remote sensing systems has continued after World War II. The second half of the 20th century saw the development of satellites, such as TIROS-I, Nimbus, and Landsat, for weather and earth observation. The information collected by the instrumentation onboard the satellites are used for civil, military, and research applications. Besides the acquisition platforms, the technology of imaging systems has also significantly advanced from using the visible spectrum to the use of near-infrared and microwave spectrums. Nowadays, satellites such as WorldView, Pleiades, GeoEye, and others are orbiting around the Earth and acquiring images of the Earth's surface at a spatial resolution as low as 0.3 m. Moreover, there are alternative image acquisition platforms, such as manned or unmanned aircraft. The remote sensing community has developed several techniques to process the images acquired and to extract useful information for various remote sensing applications.

Before the invention of photography and space technologies, people had to picture the world through artistic drawings, also called cartography, and/or text descriptions by geographers. Cartography, the science of making maps, is one of the ancient methods in which people exchanged or documented spatial information. It allowed people to understand their environment. The history of cartography dates back to the Babylonians where they created maps with topological features, such as hills and valleys, by carving them onto clay tablets [2]. Greek mapmakers produced paper maps of the known world for navigation and to represent certain areas of the earth. In addition to the Greeks, the Chinese created maps of towns, river systems, and locations, as early as fourth-century BCE, for economic reasons [2]. During the medieval times, European cartographers produced symbolic maps that are meant to represent the conception of the world at that time. These maps became outdated with the invention of Portolan Chart by the Italians in the 13th century. The age of discovery (early 15th–17th century) contributed to the creation of maps depicting new areas, such as America, explored by

Manuscript received May 20, 2018; revised September 21, 2018, December 6, 2018, and January 17, 2019; accepted January 20, 2019. Date of publication March 4, 2019; date of current version March 25, 2019. *(Corresponding author: Farid Melgani.)*

The authors are with the Department of Computer Science and Information Engineering, University of Trento, Trento 38123, Italy (e-mail: mesaybelete.bejiga@unitn.it).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTARS.2019.2895693

cartographers, merchants, and explorers. Moreover, the development of accurate cartographic techniques and the invention of tools, such as the compass, telescope, and printing press, made maps simple and accurate [2].

In addition to maps, geographers, such as Strabo, Leo Africanus, and Pausanias, depicted the places they traveled and the world known to them through writing. For instance, "The description of Greece" by Pausanias provides not only the description of topographical areas but also the description of man-made objects, such as temples, sanctuaries, and tombs. In the context of this work, converting ancient text descriptions into visual data has potential benefits to remote sensing applications. For instance, cartography in the 20th century relies on satellite images that provide detailed visual information about a given area. Hence, converting the descriptions into images can extend image-based cartography before the invention of imaging and allows to have pixel-based maps in addition to the hand-drawn maps.

The objective of this paper is to generate a pictorial summary from ancient text descriptions of a given area. This requires learning mapping between two heterogeneous data sources: text and image. Multimodal learning is a branch of machine learning that deals with building models that can relate and process information from multimodal sources [3]. Audio–visual speech processing, image captioning, multimedia retrieval, and text-to-image synthesis are examples of multimodal learning. The problem we are trying to address in this paper relates to text-to-image synthesis. In this paper, we propose to rely on generative adversarial networks (GANs) [4] to synthesize remote sensing images of the past from ancient text descriptions.

To the best of our knowledge, the idea of text-to-image synthesis is a new research topic for the remote sensing community. However, there are related works in the computer vision community worth citing. For instance, early works in mapping text descriptions into images are focused on using image retrieval techniques [5], [6]. That is, keywords or phrases from the text are used as a query to retrieve corresponding images and then the retrieved images are spatially adjusted in a way that they convey the message in the text. Another approach proposed is to use conditional generative models. In these models, text descriptions are used as a conditional information to guide the generation process. For instance, Mansimov *et al.* [7] proposed the use of generative recurrent neural networks (RNN) to convert captions into images. Yan *et al.* [8] exploited variational autoencoders (VAEs) to convert visual attributes into images. They follow a layered generation process where the foreground and background images are generated separately and composed into a single image. GANs are a new framework of generative models that formulate the modeling process as an adversarial competition between two deep networks. They have shown to generate images with better quality compared with other generative models. The success of GANs also contributed to the progress of text-to-image synthesis. More specifically, in [9]–[14], different GAN-based methods that generate image pixels by conditioning both the generator and discriminator of a GAN with text descriptions have been proposed. In addition to text, Reed *et al.* [15] control the location and content of the objects during synthesis. A recent work in the remote sensing community resorted to GANs to synthesize optical images from synthetic aperture radar (SAR) images [16], where the SAR images are used as conditioning information. This method deals with an image-to-image translation problem, and the conditioning information is directly used as an input. On the contrary, this paper deals with a text-to-image synthesis problem, which requires the conditioning information to be suitably encoded to input the network.

The present contribution is seminal and aims at opening a new research area, which is very interesting and challenging, and will attract researchers of different disciplines. From an application viewpoint, converting ancient text descriptions of geographical areas into realistic remote sensing images is particularly promising since it may impact on applications of various disciplines. Retro-remote sensing images clearly pave the way for the application and development of qualitative and quantitative information extraction approaches such as land use/cover classification and change detection methods in order to generate products with a very large temporal dimension.

In the domain of social sciences and humanities, retro-remote sensing goes with the saying that "a picture is worth a thousand words." In particular, one the disciplines that can be directly interested is landscape archeology, namely the study of how humans affected their landscape and the natural environment [17]. Researchers in this discipline analyze, assess, and interpret the formation of modern landscapes. They utilize topographic descriptions written by explorers and travelers, such as Pausanias, remote sensing images, and techniques for their research. Having the text descriptions converted to imagery could benefit researchers of the field and allow them to apply techniques for retrieving information from remote sensing images. Historical geography is another discipline that can benefit from retro-remote sensing. Indeed, researchers in this field study how geography of a certain region has changed overtime and how it impacted the events of history at a certain period [18]. Enriching their information sources with retro-remote sensing images could be valuable, in particular if the retro-remote sensing process would be fed not only with text descriptions but also with ancient hand drawings.

In summary, the main contributions of this work are as follows.

1) We propose a new field of research which has never been addressed by the remote sensing community.
2) We present a GAN-based approach to synthesize images from ancient text descriptions of geographical landscapes.
3) We also highlight open issues and possible applications of the proposed research to the remote sensing community as well as other disciplines.

This paper is organized as follows. In Section II, we present the problem formulation. Section III is dedicated to a detailed explanation of GANs. Dataset collection, experimental results and a discussion on the obtained results are presented in Section IV. Finally, we conclude by discussing open issues and proposing future research directions in Section V.

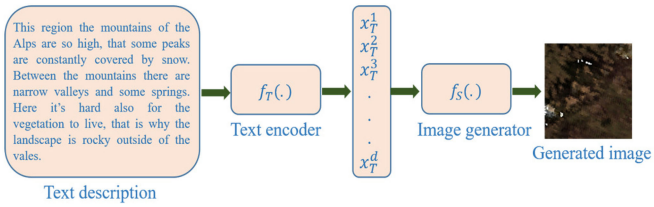


Fig. 1. General block diagram of the proposed method.

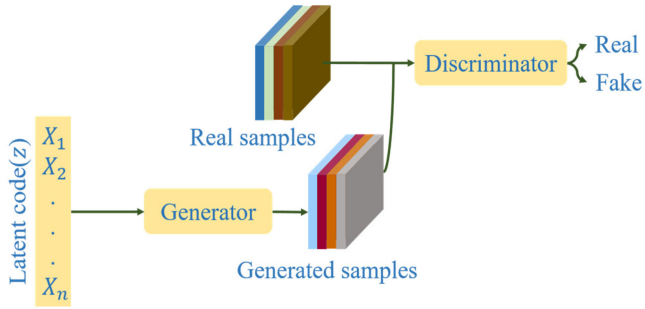


Fig. 2. The architecture of a GAN.

II. Methodology

The problem of text-to-image synthesis (see Fig. 1) combines two heterogeneous data types: text and image. The synthesis can be decomposed into two sub-problems: text representation and image generation. Text representation requires learning an efficient feature representation that has the ability to capture the required information within the sentence. Such information could be the different types of objects, the positional relationship among them, their size and color, and other useful attributes. Given a text T_n composed of a single or multiple sentences, the text representation step learns a function $f_T(.)$ that encodes the text into a feature vector $x_T \in \Re^d$, where d is the size of the vector. The field of natural language processing provides a wide range of text encoding techniques.

The second problem deals with converting the encoded text information into the pixel values of an image in a way that the information within the text is represented correctly and the image is visually appealing to a human eye. Hence, this step requires finding an appropriate model $f_S(.)$ that decodes the text into an equivalent image. An example of models that are recently being used for this purpose is generative models. Generative models focus on learning the distribution/density from which data are sampled. Models like VAEs and PixelRNNs explicitly represent the distribution over the space where the data lie. Whereas, implicit density estimation models such as Markov chain models and GANs allow to draw samples directly from the distribution (an indirect means of interaction with the distribution).

GANs are used to estimate the density p_{model} of a training set drawn from an unknown distribution p_{data}. Since introduced in 2014, GANs have been used in applications, such as domain adaptation, image-to-image translation, unsupervised feature learning, and text-to-image synthesis. The core idea of GANs is to train two networks called a generator and discriminator in an adversarial manner (see Fig. 2). The generator is part of the architecture that learns to estimate the distribution p_{model} and used to synthesize new images. The objective of the discriminator is to predict the source of input images, either they are real images or synthesized by the generator. During the training phase, generator parameters are updated based on the error from the discriminator output.

Formally, the generator of a GAN network is represented by a function G with parameters θ^G and takes a latent variable z sampled from a simple distribution p_z as an input. The discriminator is represented by a function D with parameters θ^D and takes an observed variable x as an input. Both G and D are differentiable functions with respect to their parameters, and the cost function for each network depends on the parameters of the other network. However, during training, each network has control only over its own parameters. According to this, the discriminator minimizes a cost $J^D(\theta^D, \theta^G)$ while updating only θ^D and the generator minimizes a cost $J^G(\theta^D, \theta^G)$ while updating θ^G. The solution to this optimization problem will be a (local) minimum of J^D with respect to θ^D and a (local) minimum of J^G with respect to θ^G.

In the vanilla GAN setup, the discriminator is formulated as a binary classifier with the standard cross-entropy (1) employed as a cost function. The training is performed by sampling two mini-batches of data: a mini-batch sampled from the training set and labeled as "real" and another mini-batch synthesized by the generator and labeled as "fake." In the simplest version of the game, also called the minimax (zero-sum) game, the generator minimizes negative of the discriminator loss function. Therefore, the solution to the minimax game will be the minimization of a value function (2) with respect to θ^G and maximization of the value function with respect to θ^D (3)

$$J^D\left(\theta^D, \theta^G\right) = -\frac{1}{2}\left(\mathbb{E}_{x\sim p_{\text{data}}} \log D(x) + \mathbb{E}_{z\sim p_z} \log\left(1 - D\left(G(z)\right)\right)\right) \quad (1)$$

$$V\left(\theta^D, \theta^G\right) = -J^D\left(\theta^D, \theta^G\right) \quad (2)$$

$$\theta^{G*} = \arg\min_{\theta^G}\max_{\theta^D} V\left(\theta^D, \theta^G\right). \quad (3)$$

During the initial stages of training, since the generated images are noise, the discriminator is capable of rejecting fake samples with high confidence. This causes the gradient of the generator to vanish, also called the vanishing gradient problem. In order to solve this, Goodfellow *et al.* [4] proposed a cost function (4) in which the generator maximizes the log-probability of the discriminator being mistaken rather than performing a minimization of the log-probability of the discriminator being correct

$$J^G = -\frac{1}{2}\mathbb{E}_{z\sim p_z} D\left(G(z)\right). \quad (4)$$

Mode collapse, a scenario in which the generator learns to generate only a specific class of images with different colors, themes, and/or different views, is another problem observed in the original GAN formulation. Methods such as feature matching and mini-batch discrimination [19] and combining auto-

encoder-based regularizers with an adversarial loss [20] are proposed to reduce the mode collapse problem. On the other hand, Martin *et al.* [21] introduced a new metric based on the Wasserstein-1, also called the Earth-movers (EM) distance to train GANs. Informally, if we assume distributions as a pile of masses then the EM distance measures the minimum cost required to transport the mass in order to transform one distribution into another. This cost is expressed as a product of the mass and the distance transported. Mathematically, given two distributions p and q, the Wasserstein-1 distance ($W(p, q)$) is given by

$$W\left(p,\ q\right) = \inf_{\gamma \epsilon \pi\left(p,\ q\right)} \mathbb{E}_{(x,y)\sim\gamma}\left[||x-y||\right] \tag{5}$$

where $\pi(p, q)$ is the set of all joint distributions $\gamma(x,\ y)$ whose marginal are p and q. Since the infimum in (5) in intractable, the authors proposed to use the Kantorovich–Rubinstein duality (6) instead

$$W\left(p,\ q\right) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{x\sim p}\left[f\left(x\right)\right] - \mathbb{E}_{x\sim q}\left[f\left(x\right)\right] \tag{6}$$

where the supremum is over all 1-Lipschitz functions f. Equation (6) is also intractable but it can be approximated with maximization (7) provided that the supremum is attained for some parametrized function f_w

$$W\left(p,\ q\right) = \max_{w \epsilon \mathcal{W}} \mathbb{E}_{x\sim p}\left[f_w\left(x\right)\right] - \mathbb{E}_{x\sim q}\left[f_w\left(x\right)\right]\ . \tag{7}$$

Hence, by optimizing (7), the discriminator will be able to measure the Wasserstein-1 distance between the real and fake distributions as accurate as possible. Parameters of the generator are updated using the gradient in (8):

$$\nabla W\left(p,\ q\right) = -\mathbb{E}_{z\sim p_z}\left[\nabla_{\theta^G} f_w\left(G\left(z\right)\right)\right]\ . \tag{8}$$

Weight clipping [21] and gradient penalty [22] based methods have been proposed to enforce the Lipschitz constraint. In general, the proposed metric provides a meaningful learning curve that correlates with the quality of images generated, gets rid of the mode collapse problem, and improves the stability of the training process.

The original formulation of GANs does not allow controlling the type of images being generated. For example, generating a specific class of object is not possible. In order to circumvent this problem, the Mirza and Osindero [23] proposed a conditional variant of GAN called conditional GANs (cGANs). In cGANs, the generator is tasked with generating realistic samples that satisfy the conditional information provided while the discriminators job is to predict if the generated images are realistic and also agree with the conditional information provided. The additional information used in cGANs can be class labels, parts of data, or even data from different modality. Text-to-image synthesis is a very good example of cGAN where sentences combined with a latent vector z are used to synthesize a sample that agrees with the description. Mathematically, for an additional information y, the value function in (1) is modified as follows:

$$V\left(\theta^D, \theta^G\right) = -\frac{1}{2}\left(\mathbb{E}_{x\sim p_{\text{data}}}\ \log D\left(x,\ y\right) + \mathbb{E}_{z\sim p_z} \log\left(1 - D\left(G\left(z,\ y\right)\right)\right)\right). \tag{9}$$

For the purpose of this paper, we used the Wasserstein distance to train our GAN. Hence, we modify the cost function in (7) to account for the additional information. Both the generator and discriminator will be conditioned with text information. Hence, the discriminator maximizes the distance between (text, real image) pairs and (text, fake image) pairs. In addition, following the work of Reed *et al.* [24], we add a third term on the cost function, where the discriminator maximizes the distance of (real image, wrong text) pairs from (real image, right text). This additional term will help the discriminator to identify a mismatch between an image and a text. Mathematically

$$W\left(p_{\text{data}},\ p_{\text{model}}\right) = \max_{\|f\|_L \leq 1} \begin{Bmatrix} \mathbb{E}_{x\sim p_{\text{data}}}\ \left[f\left(x,\ y^r\right)\right] \\ -\mathbb{E}_{x\sim p_{\text{data}}}\ \left[f\left(x,\ y^w\right)\right] \\ -\mathbb{E}_{x\sim p_{\text{model}}}\ \left[f\left(x,\ y^r\right)\right] \end{Bmatrix} \tag{10}$$

where y^r is the right text for an image x, and y^w is a wrong text for an image x. The generator cost function will be as follows:

$$-\mathbb{E}_{x\sim p_{\text{model}}}\left[f\left(x,\ y^r\right)\right]. \tag{11}$$

From the network architecture perspective, the generator (see Fig. 3) is a cascade of layers that perform a deconvolution operation on an input. For our work, the input will be a combination of both the noise and a text label. Whereas, the discriminator (see Fig. 4) is a cascade of convolution layers and takes images as an input sampled either from the training set or from the generator. The text label is concatenated with a fully connected (FC) layer that takes the output of the last convolution layer as an input.

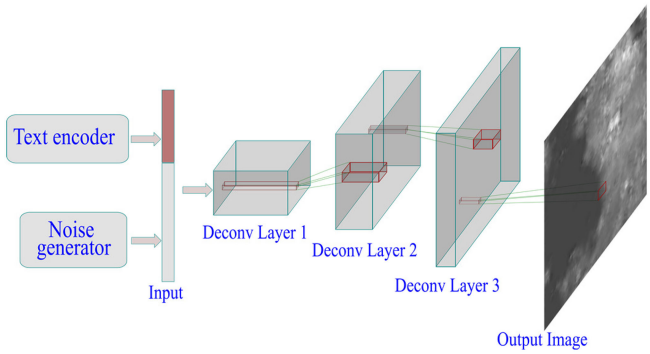


Fig. 3. Architecture of the generator.

III. Dataset Collection

A. Historical Books

As we have mentioned earlier, the objective of this paper is to synthesize remote sensing images of the past from ancient texts. In order to get these texts, we mainly referred to three books: "The Geography of Strabo: An English Translation, with

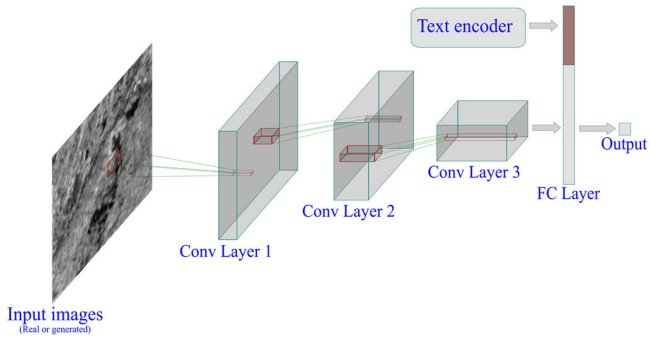


Fig. 4. Architecture of the discriminator.

Introduction and Notes" by Duane W. Roller [25], "Pausanias, Description of Greece" with an English translation by W.H.S Jones [26] (volumes VI, VII, and VIII), and "The History and Description of Africa: and of the notable things therein contained" by Leo Africanus [27] (volume I).

The first book, originally written in Greece and completed nearly 2000 years ago, was intended for the Greeks and Romans to better understand the environments they lived, moved, and/or were interested to move in. At its core, the book describes topographic, demographic, and ethnographic data about the inhibited world, starting from the southwest corner of Iberian Peninsula to India and then back west to Egypt. An excerpt of text taken from this book describing the harbor of Alexandria, Egypt is shown below:

*"Pharos is an oblong **islet**, against the mainland, making a harbour with two mouths. The shore is in the form of a **bay**, putting forward two **promontories** into the open **sea**, and the **island** is situated between them and closes the **bay**, for its length is parallel to it. The eastern **promontory** of Pharos is more toward the mainland and its **promontory** (called Lochias Promontory), and thus makes a small mouth. In addition to the narrowness of the passage in between, there are also **rocks**, some under water and others projecting from it, which at all hours make the waves strike them from the open sea rough. The extremity of the **islet** is also rock, washed all around, and there is a tower on it marvellously constructed of white stone, with many stories and named after the **island**."*

The above text contains topographic objects (highlighted in bold) such as sea, island, bay, islet, and others. It also provides information such as the relative size of objects and their relative position with the other objects. The second book, whose original version was written in the second century, used in this paper describes the topology of Attica, the Peloponnese, and central Greece mainly focusing on the sanctuaries, statues, tombs, and the legends connecting with them. It is believed that this book was mainly intended to be used as a guide-book by tourists. The following is an excerpt taken from this book:

*"There are **roads** leading from Mantineia into the rest of Arcadia, and I will go on to describe the most noteworthy objects on each of them. On the left of the **highway** leading to Tegea there is, beside the **walls of Mantineia**, a place where **horses race**, and not far from it is a **race-course**, where they celebrate the games in honour of Antinoiis. Above the race-course is **Mount Alesium**, so called from the wandering (alé) of Rhea, on which is a grove of Demeter. By the foot of the **mountain** is the **sanctuary of Horse Poseidon**, not more than six stades distant from Mantineia."*

Similarly, this text contains both man-made and natural objects (highlighted in bold) and provides information such as the spatial location of an object from other objects and the distance of an object from another object. Originally written in Arabicke and Italian by John Leo Africanus in 1550 and translated into English by John Pory in 1600, the third book provides a general description of Africa and the civilizations in the sixteenth century. An excerpt taken from this book is shown below:

*"The kingdome of Quiloa situate in nine degrees towarde the pole Antarticke, and (like the last before mentioned) taking the denomination thereof from a certaine **isle** and citie both called by the name of Quiloa; may be accounted for the third portion of the lande of Zanguebar. This **island** hath a very fresh and coole aire, and Is replenished with **trees** always greene, and with plenty of all kinde of victuals. It is situate at the mouth of the great **riuer** Coauo which springeth out of the same **lake** from whence Nilus floweth, and is called also by some Quiloa, and by others Tahiua, and runneth from the saide **lake**, eastward for the space of sixe hundred miles, till it approacheth neere the **sea**, where the streame thereof is so forcible, that at the very mouth or out-let, dispersing it selfe into two branches, it shapeth out a great **island**, to the west whereof vpon the **coast** you may behold the little **isle** and the citie arme of the **sea**."*

Similar to the examples from the others books, this excerpt also contains natural objects (highlighted in bold) such as a *sea*, *lake*, and an *island,* and gives information about the shape and spatial relationship of objects.

B. Training Set Collection

After going through the books mentioned above, we selected 43 text descriptions that contain more than 1 natural object. Since the excerpts are old, some of them contain words that are written in traditional English. Therefore, we converted these words to the equivalent modern English words. Focusing only on natural objects, the texts selected contain 27 different objects. Moreover, among the 27 objects present in the texts, we decided to focus on objects that occur with a frequency of 5 or more and also objects that can appear in low-resolution images. These objects along with their frequency of occurrence are given in Table I.

Training a GAN for our problem requires to have (image, text) pairs. However, we are considering ancient texts and the technology of satellite imaging dates back only to the 1960's. In order to solve this issue, we downloaded 12 MODIS images from Europe and Mediterranean areas and cropped images of size 100×100 in such a way that each image contains more than one objects of interest listed in Table I. An example of such images is shown in Fig. 5. Overall, we cropped 70 images. We have written the corresponding text descriptions for these training images so that to emulate the ancient stiles we saw in the books. Accordingly, we have inserted some additional information, although not directly useful for the simple text

TABLE I
TYPE OF NATURAL OBJECTS PRESENT IN THE SELECTED TEXTS AND THEIR FREQUENCY OF OCCURRENCES

Objects	# occurrences
Mountain	29
Sea/Ocean	23
Forest/Tree/Wood	15
Island	14
Grass	12
Lake	10
Coast	9
Promontory/Cape	9
Plain	8
Rock/Stone	7
Hollow/Valley	6
Gulf/Bay	5
Sand/Desert	5
Hill	5

TABLE II
EXAMPLES OF TEXT ENCODING

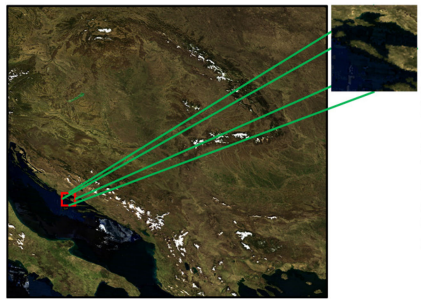
Index	Objects	Text 1	Text 2
1	Coast	1	0
2	Sea/Ocean	1	1
3	Mountain	0	0
4	Wood/Trees	0	0
5	Plain	1	0
6	Sand/Desert	0	0
7	Isle	1	1
8	Valley	0	0
9	Rock/Stone	0	0
10	Hills	0	1
11	Grass	0	0
12	Lake	0	0
13	Promontory	0	0
14	Gulf/Bay	0	1

encoding mechanism we adopted in this paper. The text on the left side of Fig. 5 is an example of the description written for the crop.

IV. EXPERIMENTAL RESULTS

A. GAN Architecture and Training Parameters

The input to the generator is a vector with a dimension of 40, of which 26 elements are noise and 14 are the text encoder outputs. During the experiments, we considered noise vector sizes ranging from 20 to 100 and we found that a vector size of 26 is better in terms of the synthesized images agreeing with the conditioned information. The input is then connected to an FC layer, which has 3136 neurons and reshaped to $7 \times 7 \times 64$. We selected the initial size of the image to be 7×7 in order to have output images with a size that is close to the true images. Using smaller initial image size will result in having more deconvolution layers and the output image size



The **coast** of mare Hadriaticum close to the city of Spalatum is irregular, starting from the **bay**, where the haven is located and almost closed by a narrow strip of land. There is also an **island** in front of the **bay**, with an elongated shape, and a **promontory** on it. Besides the first one, there are other small **islets**, not far from the **coast**. Overall, the whole zone has a low amount of vegetation and is **sandy** and covered by **hills**.

Fig. 5. An image of the Balkan Peninsula (left), a crop used for training (middle), and the corresponding text description (right).

to be much larger than desired true image size. On the other hand, using larger initial image size can reduce the number of intermediate deconvolution layer but the number parameters at the FC layer will increase significantly. The FC layer is followed by 3 deconvolution layers with 32, 16, and 8 kernels of size 5×5, respectively. The deconvolution is applied with a stride of 2 to increase the size of the image to the desired output. The output layer is also a deconvolution layer with a kernel size of 5×5, stride of 2, and 1 kernel. Except for the output layer, which uses tan*h* activation function, neurons in all other layers use a ReLu activation. It is noteworthy that the configuration reported here is among the many architecture we tried and the one that gave us better results.

The discriminator is composed of 3 convolutional layers with 8, 16, and 32 filters, respectively, and an FC layers. Each kernel has a size of 5×5 and instead of pooling, strided convolution (with a stride of 2) is used for downsampling. The inputs to the discriminator are real and fake images of size 100×100, and conditioning is done by concatenating output of the text encoder with the FC layer of dimension 100. LeakyReLu is the activation function used in the discriminator.

Training parameters are as follows.

1) Mini-batch size is set to 64.
2) Both D and G are trained iteratively. For every G, training D is trained five times.
3) We used RMSProp optimizer [28] for learning, and the learning rate is set to 0.0001.
4) Batch normalization [29] is applied to the output of every layer, except the output layers, to stabilize and speed up the training.
5) Input images are scaled in the range of $[-1, 1]$.
6) Weight clipping is applied to enforce the Lipschitz constraint with the parameter $c \in [-0.01, 0.01]$.

B. Text Encoding

In this paper, we considered a simple text encoding method. Given a text description, we look for the natural objects, more specifically the 14 class of objects listed in Table I, present in the description. The presence of an object will have a value 1 and the absence will have a value 0. Therefore, a given text description is represented by a binary vector of size 14 with each element representing the presence or absence of a specific object. For instance, given the following two text descriptions:

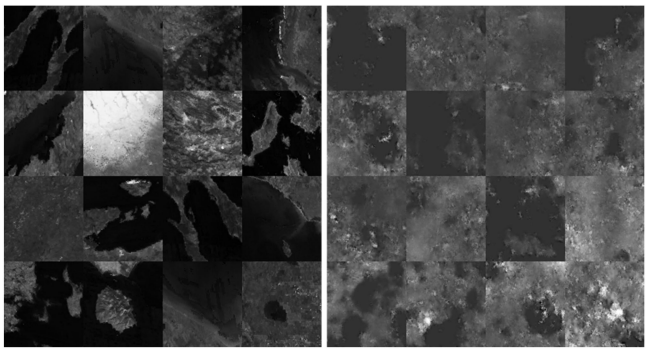


Fig. 6. Sample grayscale images from the training set (left) and GAN-generated grayscale samples (right).

In these zones in the extreme south of Italia is still located the chain of the Apennines, always extending from south to north. The ***mountains*** have low peaks, but come really close to the ***sea***, so that there is almost no beach on the ***coast*** at all. For this reason, no great harbor was built in the area.

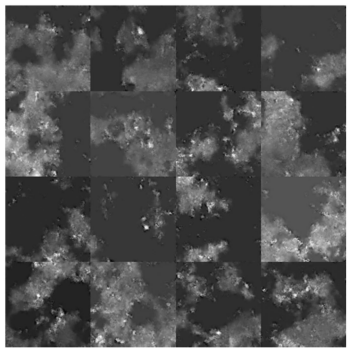


Fig. 7. Examples of grayscale images generated by the GAN conditioned with the text description shown in the left. Objects of interest are highlighted in bold-italics.

*Text 1: "Western of Lemonum there is a **plain**, which goes on until the **ocean**. This area was exploited for salt extraction and wine production after the annexation to the empire. 120 stadia far from the **coast** there is an **islet** called Yeu by the barbarians, which is uninhabited."*

*Text 2: "The Argolic **gulf** separates Arcadia to the southwest and Argolis to the north and east. It is around 300 stadia long along the **sea**. The territory is quite poor in term of vegetation and is instead mostly covered by rocks. There are many **hills** in both the zones extending on side of the **gulf** and a few little **islets** not far from the coast.*

Their corresponding binary representation is shown in Table II.

The northern part of Gallia is covered everywhere by ***plains***. The ground is flat and many rivers flow through it. There are many sparse ***forest***, but they are not dense. The people that live here are savage and mainly live from raids and by hunting.

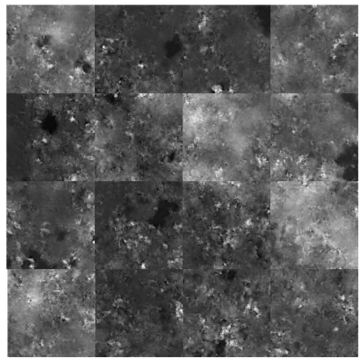


Fig. 8. Examples of grayscale images generated by the GAN conditioned with the text description shown in the left. Objects of interest are highlighted in bold-italics.

C. Results on Training Set Texts

In the context of this paper, the first experiment we conducted is to synthesize the grayscale equivalent images for the text descriptions we collected. To train the GAN, we converted the MODIS images collected into grayscale images. An example of these images is shown in Fig. 6 (left). During the training, we applied data augmentation techniques, such as flipping and rotation, to increase the number of training images to a little more than 800. After training the architecture for 2500 training epochs, we performed a visual qualitative evaluation of the generated images in different ways. The first evaluation is to condition the generator with the training set text encodings and synthesize the corresponding images. The results of this evaluation are shown in Fig. 6 (right). Additionally, we have also provided examples of training set texts along with the synthesized images in Figs. 7 and 8. Although the contrast is not as good as the training images, it is possible to see from Figs. 6–8 that the generated images are realistic and have natural shapes and textures.

In order to evaluate whether the generator is simply memorizing the training set or learning to generate images, we considered two scenarios. As a first scenario, due to the simple encoding used and multiobject nature of training labels, we synthesized samples for labels with only a single object. For instance, we conditioned the generator with equivalent labels of texts containing only mountain object and the other containing the coast object. The generated images along with the training images that contain mountain and coast in their labels are shown in Figs. 9 and 10. From Fig. 9, the generated images (right) have texture information that resembles the training set samples (left). Similarly, the generated images in Fig. 10 also contain textures of coast similar to the training examples.

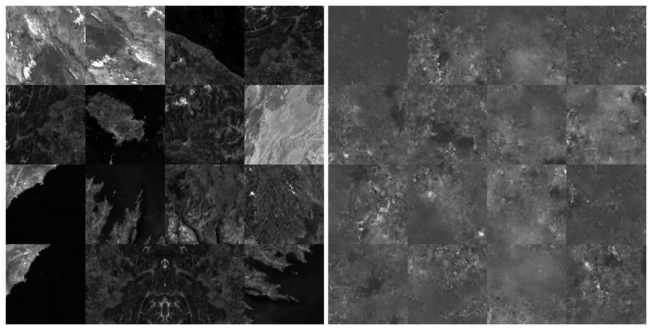


Fig. 9. Grayscale images that contain ***mountain*** label from the training set (left) and GAN-generated grayscale images (right) using ***mountain*** as only label to condition the generator.

D. Results on Ancient Texts

The second scenario we considered is to condition the generator with the ancient text descriptions (test set). Examples of the

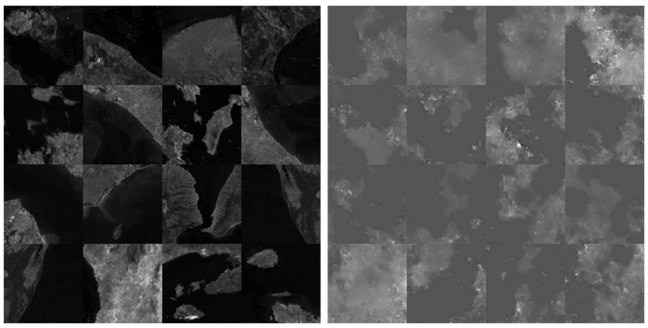


Fig. 10. Grayscale images that contain ***coast*** label from the training set (left) and GAN-generated grayscale images (right) using ***coast*** as only label to condition the generator.

Beyond the ***coast*** between the Sacred ***Promontory*** and the Pillars it is all a large ***plain***. There are many hollow: in the interior that the ***sea*** reaches, resembling moderate ravines or river channels that extend for many stadia. These are filled by the entry of the ***sea*** at the Hood tides, so that one can sail inland no less than on rivers - indeed better - for it is like sailing down rivers (as there is no resistance), since one is sent onward by the ***sea*** and the flood tide is just like the flow of a river.

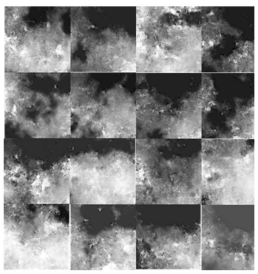


Fig. 11. Examples of images (right) synthesized by the GAN conditioned with the ancient text description (left) taken from "The geography of Strabo." Objects of interest are highlighted in bold-italics.

The entire ***coast*** of the Achaians and the others, as far as Dioskourias, and the places straight toward the south in the interior, fall at the foot of the Kaukasos. This ***mountain*** lies above both ***seas*** - the Pontic and the Kaspian - and makes a wall that extends across the isthmus that separates them. Toward the south it marks the boundary between Albania and Iberia, and toward the north, of the ***plains*** of the Sarmatians. It is well wooded with all kinds of timber, especially that used for shipbuilding. Eratosthenes says that the Kaukasos is called the Kaspios by those living there, perhaps derived from the Kaspians. There are certain arms projecting toward the south, which include the middle of Iberia and join the Armenian ***mountains*** with those called the Moschikian, and also the Skydisian and Paryadrian. These are all parts of the Tauros, which makes the southern side of Armenia, broken off in some way from it on the north and projecting as far as the Kaukasos and the ***coast*** of the Euxeinos that extends to Themiskyra from Kolchis.

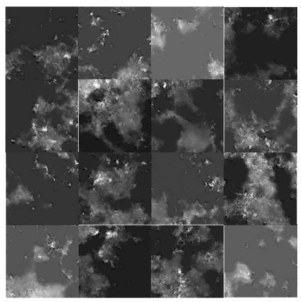


Fig. 12. Examples of images (right) synthesized by the GAN conditioned with the ancient text description (left) from taken from "The geography of Strabo." Objects of interest are highlighted in bold-italics.

generated images along with the conditioned text information are shown in Figs. 11–14. From these figures, it is possible to see that the generated images contain textures that agree with the objects described in the respective text descriptions.

E. Quantitative Results

As a quantitative evaluation, we considered conditioning the generator with ancient text labels and manually evaluating the

From the saide ***mountaines*** vnto ***mount*** Atlas there is a very spatious ***plaine*** & many little ***hillocks***. Fountaines there are in this region great store, which meeting together at one head doe send foorth most beautifull riuers and cristall streames. Betweene the foresaid ***mountaines*** and the ***plaine*** countrie is situate the ***mountaine*** of Atlas; which beginning westward vpon the ***Ocean sea***, stretcheth it selfe towards the east as farre as the borders of Aeagypt. Ouer against Atlas lieth that region of Numidia which beareth dates, being euerywhere almost ***sandie*** ground. Betweene Numidia and the land of Negros is the ***sandie*** desert of Libya situate, which containeth many ***mountainses*** also.

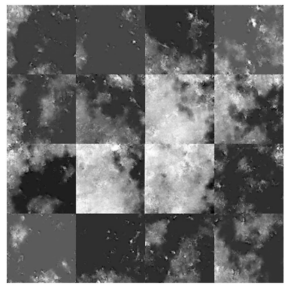


Fig. 13. Examples of images (right) synthesized by the GAN conditioned with the ancient text description (left) taken from the book by Leo Africanus. Objects of interest are highlighted in bold-italics.

The entire ***coast*** of the Achaians and the others, as far as Dioskourias, and the places straight toward the south in the interior, fall at the foot of the Kaukasos. This ***mountain*** lies above both ***seas*** - the Pontic and the Kaspian - and makes a wall that extends across the isthmus that separates them. Toward the south it marks the boundary between Albania and Iberia, and toward the north, of the ***plains*** of the Sarmatians. It is well wooded with all kinds of timber, especially that used for shipbuilding. Eratosthenes says that the Kaukasos is called the Kaspios by those living there, perhaps derived from the Kaspians. There are certain arms projecting toward the south, which include the middle of Iberia and join the Armenian ***mountains*** with those called the Moschikian, and also the Skydisian and Paryadrian. These are all parts of the Tauros, which makes the southern side of Armenia, broken off in some way from it on the north and projecting as far as the Kaukasos and the ***coast*** of the Euxeinos that extends to Themiskyra from Kolchis.

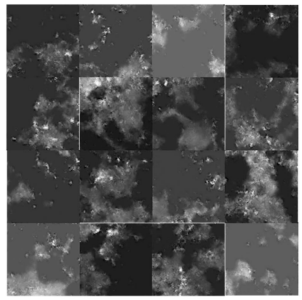


Fig. 14. Examples of images (right) synthesized by the GAN conditioned with the ancient text description (left) taken from "Pausanias, Description of Greece." Objects of interest are highlighted in bold-italics.

TABLE III
QUANTITATIVE EVALUATION OF THE GENERATED IMAGES

# Objects	Accuracy (%)
At least one	96.5
At least two	74.4
At least three	44.8
Complete disagreement	3.5

percentage of images that agree with the conditioned information. To that effect, we synthesized 16 images for each of the 43 text descriptions and evaluated the average number of images that agree with at least 1, 2, and 3 labels of the conditioned information and the average images that do not agree with the conditioned labels. From the results presented in Table III, it turns out that almost all of the synthesized images agree with at least one of the conditioned objects and only a small number of images disagree with the conditioned information. However, the accuracy in agreement decreases as the number of objects considered increases. The small number of training images could be one possible reason for that. The other reason could be that the generator network is not powerful enough to synthesize multiple objects at once.

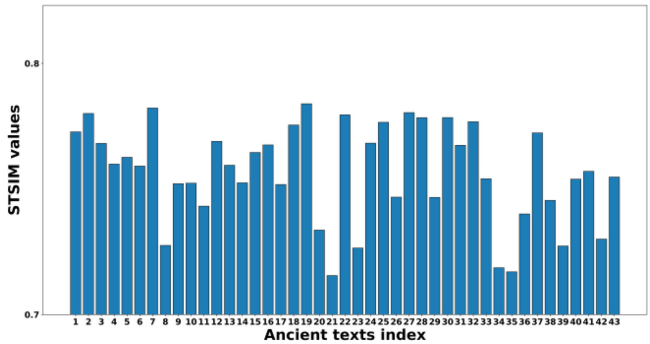


Fig. 15. Bar graph depicting the average structural texture similarity of generated images for ancient text descriptions with respect to the semantically closest training images.

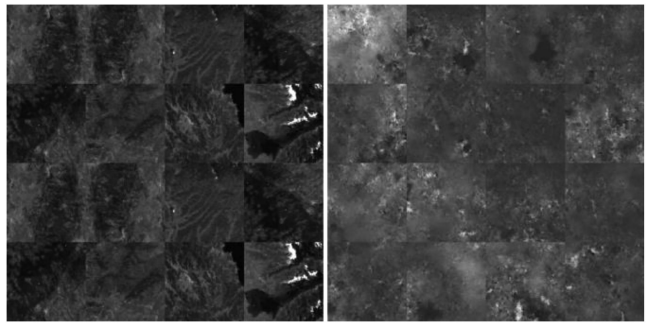


Fig. 16. Grayscale images that contain ***valley*** label from the training set (left) and GAN-generated grayscale images (right) using only ***valley*** as label to condition the generator.

In addition to this, we also applied the structural texture similarity metric (STSIM) [30] to measure the similarity between the generated and training images. That is, first we computed the Hamming distance between encoded test and training descriptions to select the most semantically closest (those with smallest Hamming distance) training images. Since the test descriptions are characterized by more than one object, we also selected training descriptions (and corresponding images) that contain more than one object to compute the distance. Once the images are selected, we computed the average STSIM value between each of the 16 images per description generated for the ancient texts and the closest training images selected. A bar graph depicting the average texture similarity of the generated images per test description is shown in Fig. 15. Overall, the generated images have a global average texture similarity of 75.6% to the semantically closest training images, which is a very encouraging result given that the synthesis task is rather complex and challenging.

F. Discussion

Although the number of training samples used to train the GAN is very small compared with the training samples used for other GAN applications, the generator was able to learn synthesizing images that have realistic textures and mostly agree with the textures of objects in the training set. In addition to this, the results also show that the generator can synthesize different but related images for a single text. However, we would like to highlight that there are scenarios where qualitative evaluation of the synthesized images for some objects was difficult. For instance, we conditioned the generator with a text label containing ***valley*** object and the textures and shapes in the generated images (see Fig. 16) agree less likely to the training set textures. One possible reason could be the shortage of enough training examples with corresponding label. Resolution of the images could also be another reason.

V. Open Issues and Future Developments

Although the topic of text-to-image synthesis is not a new topic for the computer vision community, to the best of our knowledge, it has never been explored by the remote sensing community. Moreover, the topic of retro-remote sensing is a new research field. As pioneers of this research field, we would like to highlight several issues encountered in this paper and possible research directions for the future.

A. Dataset

The first issue we would like to highlight is the size of the dataset. As we have mentioned in Section IV, the number of (image, text) pairs we used for training is very small. Since the number of parameters that need to be estimated in a GAN architecture is large, training GANs requires having a large dataset. Hence, we strongly believe that creating a large dataset and making it publicly available will significantly improve the results and advance this research area.

B. Improving the Quality of Generated Images

From the qualitative results presented in Section IV, the generated images still lack sharpness and contrast. Although the quality of images generated by GANs has been significantly improved since their introduction in 2014, this a general problem associated with GANs. Therefore, developing better algorithms is also an open area of research. For instance, in [31] a new training methodology that improves the quality of generated images has been proposed, which can be useful in the context of this paper.

C. Generating High-Resolution Images

In this paper, we used images from MODIS satellite, which have relatively low resolutions as compared with recent satellites, such as GeoEye and WorldView. Although working with low spatial resolution images is advantageous to synthesize images that cover large spatial areas, details will be missing and especially synthesizing smaller objects will be difficult. On the other hand, when working with high-resolution images, synthesizing images that cover large areas will be difficult due to the current capacity of GANs. In the current state of the art, GANs are able to synthesize images up to a size of 256×256. For instance, if we have a text description that covers an area of roughly $5\,\text{km} \times 5\,\text{km}$, it would be impossible to generate an equivalent image with a spatial resolution of $10\,\text{m}$ with a

single GAN. One possible solution could be to generate small pieces and mosaic them. Another one is first to generate a low-resolution equivalent image and then enhance the resolution. In this scenario, we can adopt a single GAN approach, similar to the work presented in [31], or resort to a cascade of GANs one for low-resolution image generation and another to enhance resolution. Overall, this is also a possible area of research within the retro-remote sensing context.

D. Color/Multispectral Image Generation

Synthesizing color/multispectral images is also another topic of research. In addition to synthesizing high-resolution images, having color images provide more information and increase the ability to discriminate between different objects. However, this will require having a deeper and more complex network architecture which requires more training examples.

E. Text Encoder

In this paper, we used a very simple encoder, as described in Section V. The shortcoming of this encoder is that it fails to encode high-level information such as the size of objects, number of objects, and the relative spatial position with each other. This may result in an image that does not fully conform to the text description, though the objects are present in the synthesized image. Alternatively, one can use off the shelf word/sentence embedding models, such as GloVe [32], Word2Vec [33], or universal sentence encoder [34], to generate better encoded forms. In general, more sophisticated text encoding mechanisms suited for this problem could be considered in future works.

F. Improving GAN Outputs With User Interaction

In their current form, GAN architectures do not have a feedback means by which they can improve the resulting output image. This kind of user/expert feedback could be particularly useful for the problem this paper is focused on. For instance, user feedback can be used as an auxiliary loss for the generator to enhance the quality of images being generated.

ACKNOWLEDGMENT

The authors would like to thank Mariette de Vos Raaijmakers (University of Trento, Italy) and Redha Attoui (University of Annaba, Algeria) for their precious advices in the selection of the ancient texts.

REFERENCES

[1] A. Khan, "History of remote sensing and GIS." [Online] Available: https://www.academia.edu/10815020/History_of_Remote_Sensing_and_GIS

[2] A. Briney, "The history of cartography," ThoughtCo, Aug. 2018. [Online] Available: https://www.thoughtco.com/the-history-of-cartography-1435696

[3] T. Baltrušaitis, C. Ahuja, and L. Morency, "Multimodal machine learning: A survey and taxonomy," in *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[4] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[5] X. Zhu, A. B. Goldberg, M. Eldawy, C. R. Dyer, and B. Strock, "A text-to-picture synthesis system for augmenting communication," in *Proc. 22nd Nat. Conf. Artif. Intell.*, 2007, pp. 1590–1595.

[6] A. B. Goldberg, J. Rosin, X. Zhu, and C. R. Dyer, "Toward text-to-picture synthesis," in *Proc. NIPS Mini-Symposia Assist. Mach. Learn. People Disabilities*, 2009.

[7] E. Mansimov, E. Parisotto, J. L. Ba, and R. Salakhutdinov, "Generating images from captions with attention," in *Proc. Int. Conf. Learn. Represent.*, 2016, arXiv: 1511.02793.

[8] X. Yan, J. Yang, K. Sohn, and H. Lee, "Attribute2Image: Conditional image generation from visual attributes," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 776–791.

[9] H. Zhang *et al.*, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5908–5916.

[10] H. Zhang *et al.*, "StackGAN++: Realistic image synthesis with stacked generative adversarial networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018, to be published, doi: 10.1109/TPAMI.2018.2856256.

[11] A. Dash, J. C. B. Gamboa, S. Ahmed, M. Liwicki, and M. Z. Afzal, "TAC-GAN—Text conditioned auxiliary classifier generative adversarial network," Mar. 2017, arXiv: 1703.06412.

[12] H. Dong, S. Yu, C. Wu, and Y. Guo, "Semantic image synthesis via adversarial learning," in *Proc. IEEE Int. Conf. Comput. Vis, Venice*, 2017, pp. 5707–5715.

[13] A. Nguyen, J. Clune, Y. Bengio, A. Dosovitskiy, and J. Yosinski, "Plug & play generative networks: conditional iterative generation of images in latent space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, 2017, pp. 3510–3520.

[14] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1060–1069.

[15] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, "Learning what and where to draw," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 217–225.

[16] P. Wang and V. M. Patel, "Generating high quality visible images from SAR images using CNNs," in *Proc. IEEE Radar Conf.*, 2018, pp. 570–575.

[17] S. H. Parcak, "GIS, remote sensing, and landscape archaeology," Oxford Handbooks, Mar. 2017. [Online] Available: http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199935413.001.0001/oxfordhb-9780199935413-e-112

[18] R. A. Butlin and B. Graham, *Historical Geography: Through the Gates of Space and Time*. London, U.K.: Edward Arnold, 1993.

[19] T. Salimans *et al.*, "Improved techniques for training GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2234–2242.

[20] T. Che, Y. Li, A. Jacob, Y. Bengio, and W. Li, "Mode regularized generative adversarial networks," in *Proc. Int. Conf. Learn. Represent.*, Feb. 2016, arXiv: 1612.02136.

[21] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.

[22] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5767–5777.

[23] M. Mirza and S. Osindero, "Conditional generative adversarial nets," Nov. 2014, arXiv: 1411.1784.

[24] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1060–1069.

[25] D. W. Roller, *The Geography of Strabo: An English Translation, With Introduction and Notes*. Cambridge, U.K.: Cambridge Univ. Press, 2014.

[26] W. H. S. Jones and H. A. Ormerod, *Pausanias: Description of Greece, vol. II, Books 3-5 (Laconia, Messenia, Elis 1) (Loeb Classical Library No. 188)*. New York, NY, USA: William Heinemann/GP Putnam's Sons, 1926.

[27] A. Leo, J. Pory, and R. Brown, *The History and Description of Africa*. London, U.K.: Printed Hakluyt Soc., 1896.

[28] G. Hinton, N. Srivastava, and K. Swersky, *Neural networks for machine learning. Lecture 6a—Overview of mini-batch gradient descent*, Coursera Lecture slides, 2012.

[29] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[30] J. Zujovic, T. N. Pappas, and D. L. Neuhoff, "Structural similarity metrics for texture analysis and retrieval," in *Proc. 16th IEEE Int. Conf. Image Process.*, 2009, pp. 2225–2228.

[31] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. Int. Conf. Learn. Represent.*, Feb. 2018, arXiv: 1710.10196.
[32] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Nat. Lang. Process.*, Doha, Qatar, 2014, pp. 1532–1543.
[33] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. Int. Conf. Learn. Represent.*, Feb. 2013, arXiv: 1301.3781.
[34] D. Cer *et al.*, "Universal sentence encoder," Mar. 2018, arXiv: 1803.11175.

Mesay Belete Bejiga received the M.Sc. degree in telecommunications engineering from the University of Trento, Trento, Italy. He is currently working toward the Ph.D. degree in signal processing and pattern recognition at the ICT Doctoral School, University of Trento.

His research interests include machine learning and image processing techniques applied to remote sensing problems.

Farid Melgani (M'04–SM'06–F'16) received the State Engineer degree in electronics from the University of Batna, Batna, Algeria, in 1994, the M.Sc. degree in electrical engineering from the University of Baghdad, Baghdad, Iraq, in 1999, and the Ph.D. degree in electronic and computer engineering from the University of Genoa, Genoa, Italy, in 2003.

From 1999 to 2002, he cooperated with the Signal Processing and Telecommunications Group, Department of Biophysical and Electronic Engineering, University of Genoa. Since 2002, he has been an Assistant Professor and then an Associate Professor of telecommunications at the University of Trento, Trento, Italy, where he has taught pattern recognition, machine learning, radar remote-sensing systems, and digital transmission. He is the Head of the Signal Processing and Recognition Laboratory, Department of Information Engineering and Computer Science, University of Trento. His research interests include the areas of remote sensing, signal/image processing, pattern recognition, machine learning, and computer vision. He has coauthored more than 200 scientific publications and is a referee for numerous international journals.

Dr. Melgani has served on the scientific committees of several international conferences and is currently an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, *International Journal of Remote Sensing*, and IEEE JOURNAL ON MINIATURIZATION FOR AIR AND SPACE SYSTEMS.

Antonio Vascotto received the bachelor's degree in electronics and telecommunications engineering and the M.Sc. degree in telecommunications from the University of Trento, Trento, Italy, in 2015 and 2018, respectively. He is currently working with one of the world leaders in the field of energy management, focusing on automation and digital transformation. His main research interests, during his studies, include image processing and machine learning.