Synthesis of Satellite-Like Urban Images From Historical Maps Using Conditional GAN

Henrique J. A. Andrade and Bruno J. T. Fernandes, *Member, IEEE*

***Abstract*— One method for encouraging the public interest in the use of historical maps as a source of reliable knowledge is to represent them in a more familiar aspect, such as the style of the current-day popular application Google Maps' satellite view. We present a method for synthesizing satellite-images from historical maps, translating their visuals using conditional generative adversarial networks (conditional GANs). We discuss a typical representation of these dated documents to allow such translations. We observe how the semantics involved in the process influence the outcomes. Finally, we discuss the effective result of bringing the past to a familiar look for the viewer.**

***Index Terms*— Neural networks, urban areas.**

I. Introduction

CARTOGRAPHIC documents are complex texts with which humans organize and communicate their spatial knowledge of the world. They are culturally rich and socially significant documents. Their study is as revealing and as rewarding as that of any other work of art, literature, or science [1]. As maps are designed by one society to represent and guide its interests, they serve as strategical documents that yield political representation to those communities there depicted. Historical maps are an irreplaceable primary source of geographical and political information from the past [2]. The geographical documentation of some community in the past reassures the importance of its present existence, likewise, the omission of communities in these documents erases them from history. Furthermore, maps serve as tools for controlling historical hegemony [3].

The understanding of old maps' complex importance and implications requires us to bring them to the center of historical inquiries. Contrasting with the current-day tools for mapping regions, such as the use of satellite imagery and Lidar scanning, historical maps are products of manual labor, as seen in the example of Fig. 1. The art of elaborating these documents traces back to almost a millennium ago. As a consequence of the evolving drawing techniques from primitive and often distorted spatial representations to documents accurate to the centimeter, information presented in old maps might not be easily read by the ordinary observer. Additionally, historical spatial data are not as easily accessible as modern maps are. These documents are sparsely available on the *Internet* or remotely found in museums and few public collections.

Fig. 1. Cutout of a map of Recife, Brazil, in 1808. It shows the 19th century urban development of the city and its environment with plenty of water bodies [5].

As the reconstruction of maps depends on the available data of a desired region, literature discusses computational methods, either backtracking models, retrospective interpolation, or Markov chain models [4]. Nevertheless, the task of synthesizing a satellite-like image as a more up-to-date version of a map imposes a problem: a texture-based approach needs contextual representation of some dated urban architecture layouts, including its long-modified natural surroundings. Hence, solutions need to incorporate some context in their translation.

In this work, we discuss an approach for synthesizing satellite images out of historical cartographic documents, presenting a more familiar look for the observer. Using an image-to-image translation method, we propose adding an intermediate step, which consists of a common visual representation between the two endpoints of the framework.

Furthermore, we discuss the role of supporting semantics for the given task and the use of multidisciplinary support from the literature. Finally, we compare different approaches and results, and we debate the importance of architectural and geological-precise semantics for more accurate historical representations.

II. Historical Map Synthesis

This work presents an architecture that involves the creation of sets of label maps from the inputs, the translation of those label maps, and the combination of results, as visualized in Fig. 2. The output is a synthesized satellite image that combines the texture of the inputs to reproduce visuals that allow the observer to imagine that period of time. For a more concise output, the input textures need to represent visuals approaching the context from the old map in analysis.

Manuscript received June 15, 2020; revised August 18, 2020; accepted September 3, 2020. This work was supported in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil, under Grant 001 and in part by Brazilian Agencies Fundação de Amparo à Ciência e Tecnologia de Pernambuco (FACEPE) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). *(Corresponding author: Bruno J. T. Fernandes.)*

The authors are with the Escola Politécnica da Universidade de Pernambuco, Recife 50720-001, Brazil (e-mail: hjaa@ecomp.poli.br; bjtf@ecomp.poli.br).

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

Digital Object Identifier 10.1109/LGRS.2020.3023170

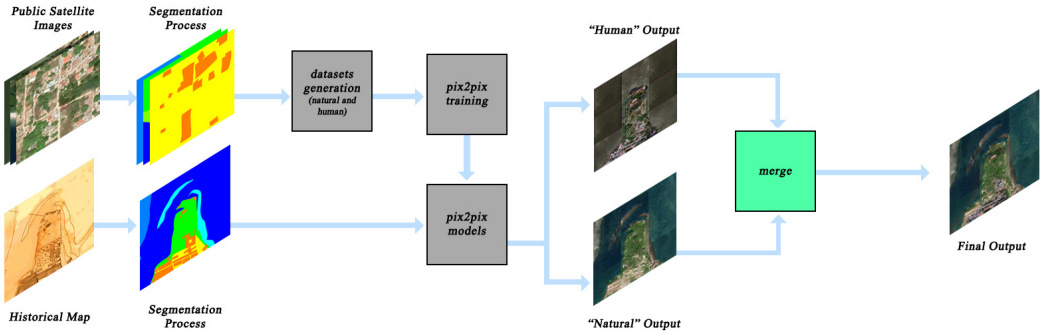

Fig. 2. We first generate label maps with different information of interest, then we select satellite images covering the labels of the map in analysis. The satellite images and their respective label maps are the training inputs for the *pix2pix* architecture. Finally, regarding the label of interest, we merge both of the results to compose the final image.

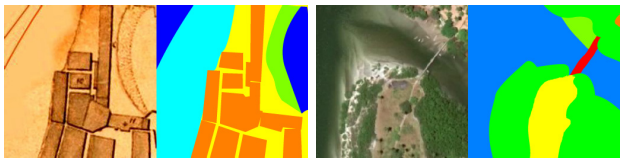

Fig. 3. Examples of label maps generated from the map, and satellite images, used as inputs for this work.

Fig. 4. Samples of satellite images extracted from more rural settings and natural landscapes.

A. Historical Map Segmentation

In image-to-image translation, label maps are often used as context-driven inputs. Wang *et al.* [6], [7] and Zhu *et al.* [8] use segmented labels to generate images in different applications. We consider context to be the localization of textures to reproduce visuals from the time period of the analyzed map. Context brings semantics to our architecture which yields a more convincing synthesized image for the observer. In other words, gathering our system's inputs requires supporting data that justifies the use of such inputs.

It was only during the 20th century that the map production shifted from handcrafted work to be created by computers. Current-day digital map platforms provide many manipulation aids such as dynamic text labels and colors, facilitating interpretation for the reader. Conversely, printed maps only counted with a legend and other reading aids found within the actual document. Supporting visual aids in printed maps are a rich source of context as they describe details about the terrain or the types of land occupation, such as residential, religious, urban equipment, industrial, ports, and others. In addition, architectural and anthropological documentation of the site can also aid as a supporting source for context [4]. Samples of the label maps generated for this work are displayed in Fig. 3.

In order to approximate the input's context to the desired output, textures need to resemble that of the time's urban configuration. In most cases, current-day satellite images may not reflect the setting of that region in the past, especially regarding city maps. To overcome changes in urban landscapes throughout time, a suggestion would be extracting textures from small towns around the region in analysis (Fig. 4). With supporting references, satellite images from other regions of the globe can also serve as inputs.

B. Pix2pix Method

Proposed by Isola *et al.* [9], *pix2pix* is a conditional generative adversarial network (cGAN), a framework for image-to-image translation that is not application-specific. One of the many purposes discussed by its authors is the translation between top-down street maps to satellite-like images. Therefore, as we have our segmentation maps, we can perform the image-to-image translation, from label maps to satellite-like images, using the *pix2pix* method.

The *pix2pix* architecture consists of a generative network G and a discriminator network D. We use G to learn how to translate segmentation maps into satellite-like images, while the discriminator D judges the results classifying them as real or computer-generated images. As the framework is a supervised method, inputs are provided in pairs of the expected satellite-like image and its label map.

A GAN will learn to map from a random noise vector z to an image y, $G : z \rightarrow y$ [10]. In comparison, a conditional GAN will also consider an image x as input to output y, $G : x, z \rightarrow y$. As for this work, G is then trained to produce satellite-image outputs not distinguishable from the set of inputs. The task of distinguishing between elements is then given to the network D. The objective of a conditional GAN can be expressed as

$$\mathcal{L}_{\text{cGAN}}(G, D) = \mathbb{E}_{x,y}[\log D(x, y)] + \mathbb{E}_{x,y}[\log (1 - D(x, G(x, z))]. \quad (1)$$

The authors also discuss a comparison to an unconditional variant where the discriminator network D does not observe x

$$\mathcal{L}_{\text{GAN}}(G, D) = \mathbb{E}_{y}[\log D(y)] + \mathbb{E}_{x,y}[\log (1 - D(G(x, z))]. \quad (2)$$

As proposed by the *pix2pix* method, the generative network G's objective function uses a $L1$ distance [9], and it is tasked to fool the discriminator D, while remaining near to the ground truth output

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[||y - G(x, z)||_1]. \tag{3}$$

While G aims to reduce its loss function and D aims to raise it, the *pix2pix*'s final objective can be written as

$$G^* = \arg\min_G \max_D \mathcal{L}_{\text{cGAN}}(G, D) + \lambda \mathcal{L}_{L1}(G) \tag{4}$$

where

x	the segmented label map input
y	the satellite image output
z	a random noise vector
D	the discriminative network
G	the generative network
λ	constant to control artifacts on the cGAN result.

The generative network G adopts a U-Net architecture [11], while the discriminator D adopts a custom patch-based fully convolutional network. Inputs are colored images, each measuring 256 $\times$ 256 pixels, concatenated side by side. We use the texture maps and their labels to train and test a model, then we finally evaluate it over the historical label maps.

C. Merging Results

Different inputs may be more appropriate for a desired representation. Textures depicting urban settings will likely yield a more realistic urban-like output, as natural landscapes' textures will probably represent nature better. Additionally, depending on the segmentation methods or the individual performing it, if manually, we may also come up with different label maps. As each divergent input sets may be better at representing some specific classes, our solution proposes a merge of different training results. We first generate different label maps, we then apply *pix2pix* to each of the results, and finally, we merge the translated label maps into one final image.

To perform a merge, we specify which label each translated image was trained to focus. The result is a composition of regions of interest from the inputs, as illustrated in Fig. 2. We evaluate each of the pixels p from the generated images using their respective networks G_n and compose the final image y as described in

$$y(p) = \begin{cases} G_1(s_1)(p), & \text{if } p \text{ is part of label of interest} \\ G_2(s_2)(p), & \text{otherwise} \end{cases} \tag{5}$$

where

p	pixel from image
G_n	the generative network trained for each of the experiments
s_n	the input image for each of the experiments
y	the output image.

The inputs and other implementation information, as well as our label of interest, are described in Section III.

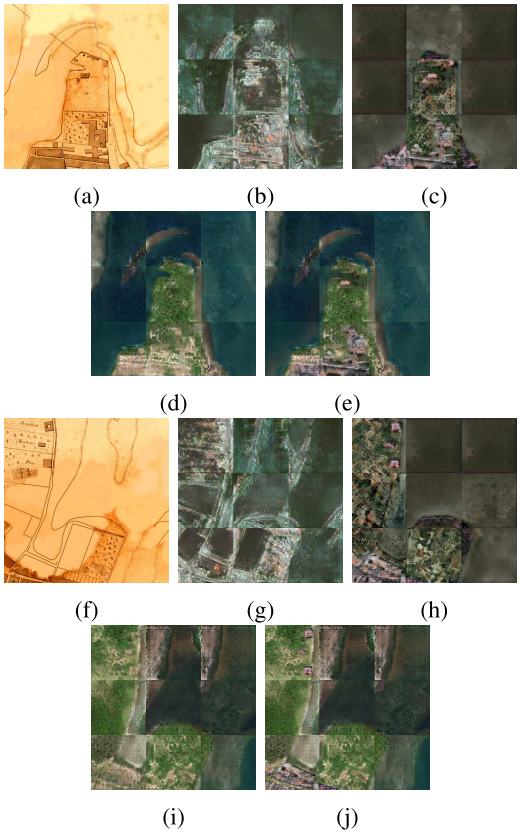

Fig. 5. Comparison between the original (a) and (f) and generated results. In (b) and (g), we display the result of state-of-the-art *pix2pix*, without the use of label maps. Images (c), (d), (h), and (i) are the results of our first and second experiments. Finally, we display the merge of experiments' results in (e) and (j).

III. Results

For this work, we used a map from 1808 Recife, Brazil (Fig. 1). Satellite images were obtained from *Google Earth* and were segmented by hand. We executed two experiments with different sets of inputs (Fig. 3). The first set contains images of current-day Recife, hence, a more predominantly city-style imagery. Thus, the urban setting is our label of interest. In other words, we use the pixels labeled as urban from the results of this network. The second set covers more rural areas and natural landscapes from areas around Recife. We trained *pix2pix* for 300 epochs each experiment using *Google Research Colab* tools in Python, where each epoch took an average of 19.29 s to execute. We applied $\lambda = 100$ as suggested by Isola *et al.* [9]. When trained, our model takes 0.27 s to output an image on average.

We compare our results with the output of state-of-the-art *pix2pix* in Fig. 5. For means of comparison, we extracted the edges from *Google Maps* patches, and pairing with satellite images of the same area, we used them as inputs for training. We then apply the trained model to the historical map image with its edges extracted as well, reproducing the *pix2pix* original experiment of translating maps to satellite images. When trained, the *pix2pix* generative network has significant power of generalization, however, without the adequate context, the output fails to reproduce trustworthy visuals.

Fig. 6. Our synthesized satellite image of 1808 Recife and the city present-day actual satellite picture. One of our goals is bringing awareness for the landscape changes.

Comparing those results with the contextualized outputs, our first experiment [Fig. 5(c) and (h)], one can observe textures from a particularly modern urban setting. Other visual elements resemble settings yet to exist and do not reflect the map in analysis' period of time. In parallel, the second experiment results display a focus on synthesized natural aspects [Fig. 5(d) and (i)]. As the second set of label maps does not contain dense urban settings, buildings were poorly represented when compared with the results from the first experiment.

A combination of the results of both experiments shows a closer representation to the description of the object in analysis. Merging results, considering the pixels labeled as urban settings from the first experiment, and applying them to the second experiment result yields a trustworthy output, more recognizable than the previous results for the observer [Fig. 5(e) and (j)].

Even with fewer segmentation classes and training samples, the first experiment resulted in a better reproduction of urban visual patterns. However, the results for the second experiment yield a closer visual to what Recife must have resembled in that period, thus bringing about a better localization.

IV. Conclusion

Our results allude to a close visual to what Recife must have resembled in the past. This work casts a light over this field of study, it employs computer science to a social application, where it may serve as an exceptionally necessary tool for historians and other scholars. The reconstruction of such documents impacts postcolonial debates and the civil engineering industry. Presenting a version of one's surroundings from the past may assist in the understanding of changes that influenced society to its current state. It may also point to where it should go, either to repair errors from the past or to avoid errors in the future. In Fig. 6, we can observe the changes in urban settings between our synthesized image from 1808 and the satellite image from 2020.

Advancing this work, next steps would involve generating a higher definition for the resulting images, either by improving the segmentation method or the translation method. Furthermore, a deeper examination of the satellite imagery extraction, corroborated by specialized literature, has been proved to yield trustworthy synthesized images. Moreover, the number of classes to represent different types of land-use may also be reviewed, however, because of the versatility of the *pix2pix* method, same-class pixels arrangements are generalized in a satisfactory way.

References

[1] J. Dym and K. Offen, *Mapping Latin America: A Cartographic Reader*. Chicago, IL, USA: Univ. Chicago Press, 2011.

[2] Y. Chiang, W. Duan, S. Leyk, J. Uhl, and C. Knoblock, *Using Historical Maps in Scientific Studies: Applications, Challenges, and Best Practices*. Springer, 2020, doi: 10.1007/978-3-319-66908-3.

[3] V. B. Tuncay, "Reflexiones sobre el uso del material cartográfico como herramienta pedagógica en América latina: Una función marginalizada ante la función estratégico-legal," *Apuntes. Revista de estudios sobre patrimonio cultural*, vol. 26, no. 1, pp. 1–10, May 2014.

[4] Y. Yang, S. Zhang, J. Yang, L. Chang, K. Bu, and X. Xing, "A review of historical reconstruction methods of land use/land cover," *J. Geographical Sci.*, vol. 24, no. 4, pp. 746–766, Aug. 2014.

[5] G. Ferrez, "Plano do porto e praça de Pernambuco por José Fernandes Portugal. Piloto que serviu n'armada Real. Anno 1808," Fundação Nacional Pró-Memória, Fundação do Patrimônio Histórico e Artístico de Pernambuco, Recife, Brazil, Tech. Rep., 1984.

[6] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.

[7] X. Wang, K. Yu, C. Dong, and C. C. Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 606–615.

[8] S. Zhu, S. Fidler, R. Urtasun, D. Lin, and C. C. Loy, "Be your own prada: Fashion synthesis with structural coherence," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1680–1688.

[9] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

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

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)* Springer, 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.