Procedural Terrain Generation Using Generative Adversarial Networks

1st Georgios Voulgaris
Aristotle University of Thessaloniki
Department of Informatics
Thessaloniki, Greece
voulgeor@csd.auth.gr

2nd Ioannis Mademlis
Aristotle University of Thessaloniki
Department of Informatics
Thessaloniki, Greece
imademlis@csd.auth.gr

3rd Ioannis Pitas
Aristotle University of Thessaloniki
Department of Informatics
Thessaloniki, Greece
pitas@csd.auth.gr

***Abstract*—Synthetic terrain realism is critical in VR applications based on computer graphics (e.g., games, simulations). Although fast procedural algorithms for automated terrain generation do exist, they still require human effort. This paper proposes a novel approach to procedural terrain generation, relying on Generative Adversarial Networks (GANs). The neural model is trained using terrestrial Points-of-Interest (PoIs, described by their geodesic coordinates/altitude) and publicly available corresponding satellite images. After training is complete, the GAN can be employed for deriving realistic terrain images on-the-fly, by merely forwarding through it a rough 2D scatter plot of desired PoIs in image form (so-called "altitude image"). We demonstrate that such a GAN is able to translate this rough, quickly produced sketch into an actual photorealistic terrain image. Additionally, we describe a strategy for enhancing the visual diversity of trained model synthetic output images, by tweaking input altitude image orientation during GAN training. Finally, we perform an objective and a subjective evaluation of the proposed method. Results validate the latter's ability to rapidly create life-like terrain images from minimal input data.**

***Index Terms*—Artificial Terrain, Generative Adversarial Networks, Deep Learning, Procedural Content Generation.**

I. Introduction

Manually-created virtual terrains are still superior in quality than ones derived with automated means, at the cost of significant labour and time expenses. The complexity of the real world (rocks, grass, trees, mountains) renders the creation of plausible, original terrain content still a challenging task. This issue can be bypassed using Procedural Content Generation (PCG), i.e., a set of methods for (semi-)automatically creating new content for 2D/3D graphics on-the-fly, and thus replacing the artistic part of content generation with a choice of tweakable parameters and random elements. PCG algorithms can be used for on-the-fly creating 2D *terrain images* that encode 3D characteristics (e.g., altitude); this terrain image can then be transformed into a 3D terrain mesh at a final post-processing step.

Typical noise-based terrain generators (e.g., Worley [1], simplex [2], Perlin [2], value [3] or diamond-square [4]) suffer with regard to memory/computational requirements and/or output quality. More recent PCG approaches that have been applied for terrain generation, such as Software Agents [5], Erosion Modeling [6] and Evolutionary Algorithms [7] also typically require significant manual post-processing (e.g., applying an image overlay to achieve a realistic look) and/or extensive manual parameter tuning.

Thus, Deep Neural Networks (DNNs) such as Generative Adversarial Networks (GANs) [8] have been alternatively explored for visual content generation. In [9] a GAN-based method is presented for multi-scale terrain texturing with reduced tiling artifacts. It involves training a GAN to upsample and texture map a low-resolution terrain input. Thus, during the inference stage, low-resolution terrain images can be translated on-the-fly to high-resolution ones; thus the terrain is needed upfront as input to be up-scaled. Other GAN-based methods [10] [11] create mountain-like 3D terrains, using information extracted from training height map data. Acquiring height maps is not trivial, while the generated results need to be heavily post-processed, since they are missing textures and realistic visual features (e.g., grass, rivers, forests, etc.).

In comparison, this paper presents a novel GAN-based method for procedural terrain generation with significantly more relaxed input data requirements (very loose constraints are only imposed upon the input data) and a higher diversity of terrain results. We call this proposed method *GAN-terrain*. Unlike other GAN-based terrain generation methods, it does not require sophisticated input data types (e.g., height maps). Thus, after training, it only incurs minimal manual supervision, since its required input simply consists of easily constructed (in a matter of seconds), rough 2D scatter plots of desired Points-of-Interest (PoIs) in image form; we call such a plot an "altitude image". The output is a 2D textured terrain resembling a satellite image, with colour encoding height and/or geomorphological properties (e.g., snow, water-body, forest, etc.), so that it can then be trivially post-processed and converted into a semantically annotated 3D terrain mesh. During training, the model learns to extract altitude/spatial information from colour density/distances of input PoIs.

GANs can easily learn complex real-word semantic content, like mountains, sea, deserts, islands, or flora, in a way that follows natural spatial alignment constraints (e.g., no jungles depicted in frozen Arctic regions, no rivers flowing uphill, etc.). However, simply training a GAN on a large set of

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911 (AI4Media).

ground-truth terrain images does not guarantee that the Generator will learn to produce complex content that obeys similar restrictions. Therefore, we opted for an Image-to-Image Translation GAN, training it using geographic coordinates and altitude information from a dataset of neighbouring landmarks, paired with the corresponding satellite image of their region.

The main advantage of GAN-terrain lies in its novel input strategy, that simplifies the actual use of the deployed DNN model on the field: new inputs for the trained network, i.e., novel altitude images at the inference stage, can be trivially created in a matter of seconds with any image processing software. In fact, although the training/evaluation dataset for this paper was constructed using real geographic data, we have successfully tested the trained GAN-terrain model with arbitrary input images; the Generator still predicts relatively realistic terrain images.

The only existing methods partly similar to GAN-terrain are [12] and [13]. However, the first one also requires height maps for training, while both of them rely on unconditional GANs for 2D terrain image/texture generation. In contrast, GAN-terrain does not require height maps and is built upon the Image-to-Image Translation framework for increased robustness.

II. GAN-Terrain Method

Generative Adversarial Network (GAN) are employed as the primary tool for completing the procedural terrain generation task. GANs are composed of two sub-networks being trained jointly, namely a Generator (G) and a Discriminator (D). After training, only the Generator is typically retained for content generation purposes. In this paper, the *conditional GAN* variant for Image-to-Image Translation tasks is employed [14]. GAN theory and training is briefly presented below (details in [8] [14]).

A. Generative Adversarial Networks

In an image synthesis scenario, GANs are generative models that learn a mapping $G : \mathbf{z} \mapsto \mathbf{Y}$ from a random noise vector $\mathbf{z} \in \mathbb{R}^n$ to output image $\mathbf{Y} \in \mathbb{R}^{k \times l}$ or a tensor $\mathbf{Y} \in \mathbb{R}^{k \times l \times m}$. In contrast, conditional GANs learn a mapping $G : \{\mathbf{X}, \mathbf{z}\} \mapsto \mathbf{Y}$ from observed input image $\mathbf{X} \in \mathbb{R}^{p \times r}$ and random noise vector $\mathbf{z}$, to $\mathbf{Y}$.

The Generator G is trained to produce outputs that cannot be distinguished from "real" images by an adversarially trained Discriminator D, which gradually learns to discern the synthetically generated images from real ones. The objective of a conditional GAN can be expressed as:

$$\begin{aligned} L_{cGAN}(G, D) = \mathbb{E}_{\mathbf{X},\mathbf{Y}}\left[logD\left(X, Y\right)\right] + \\ + a\mathbb{E}_{X,z}\left[log\left(1 - D\left(\mathbf{X}, G\left(\mathbf{X}, \mathbf{z}\right)\right)\right)\right], \end{aligned}$$

where G tries to minimize this objective against an adversary D that tries to maximize it:

$$G^* = arg \min_G \max_D L_{cGAN}(G, D). \quad (1)$$

In the unconditional variant, where the Discriminator does not observe $\mathbf{X}$, it holds that:

$$L_{cGAN}(G, D) = \mathbb{E}_Y\left[logD\left(\mathbf{Y}\right)\right] + \mathbb{E}_{X,z}\left[log\left(1 - D\left(G\left(\mathbf{X}, \mathbf{z}\right)\right)\right)\right]. \quad (2)$$

It is best practice to augment the GAN objective with a more traditional loss, such as L_1 or L_2 norm. Although the Discriminator's job remains unchanged, the Generator is additionally constrained to stay near the corresponding ground-truth output as follows:

$$L(G) = E_{X,Y,z}\left[\|\mathbf{Y} - G\left(\mathbf{X}, \mathbf{z}\right)\|\right]. \quad (3)$$

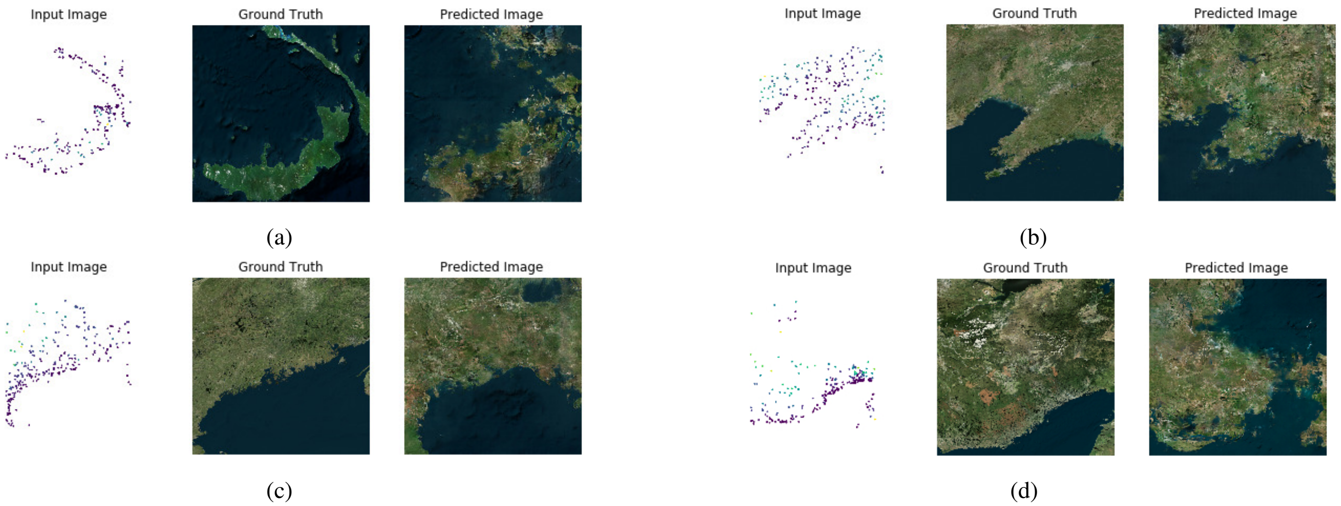
The overall training objective is:

$$G^* = arg \min_G \max_D L_{cGAN}(G, D) + \lambda L(G). \quad (4)$$

B. GAN-Terrain

The proposed *GAN-terrain* method consists in training a conditional GAN for image synthesis so that it learns to map rough 2D Point-of-Interest (PoI) scatter maps (so-called *altitude images*) into realistic satellite terrain images containing geomorphological details. In the inference/deployment stage, after training has been completed, a similar altitude image can be easily crafted at minimal labour and time expense (within seconds), in order to be fed to the trained model as observed input image $\mathbf{X}$. The corresponding model output $\mathbf{Y}$ will be a procedurally generated 2D terrain image with rich, color-coded geomorphology that typically does not violate spatial intuitions.

To train the desired conditional GAN model under this framework, we initially collect a set of N earth surface PoIs $\mathbf{p}_i = [\lambda_i, \phi_i, R_i]^T \in \mathbb{R}^3$, $i = 1, ..., N$, composed of longitude λ_i, latitude ϕ_i and altitude R_i components. The altitude is rescaled and quantized to integer interval $[0, 253]$, assuming the height of the mount Everest (8.848m) is the maximum possible value. These N vectors can be grouped into *geographic patches*, i.e., rectangle-shaped earth regions defined from 4 PoIs. Subsequently, this set is uniformly sampled to select a set of M geographic patches, so that most earth region terrain variations are represented on the training dataset. Such a representation of all earth terrain variations is essential for high-quality, diverse content generation. Finally, for each of the M geographic patches, we collect a random number of PoIs falling geographically within it, as well as a satellite image of the patch. Patch PoIs $\mathbf{p}_{ji}$,$i = 1, ...N$ are employed to construct a 2D altitude image (λ_j, ϕ_j), of patch $j = 1, ...M$ where the horizontal/vertical coordinate corresponds to PoI latitude/longitude$(\lambda_{ji}, \phi_{ji})$, respectively, while the luminance of each point encodes PoI normalized altitude R_{ji}. Such altitude images are very sparse, since typically we sample only few Earth surface points p_{ji},$j = 1, ...M$ per patch. All other altitude image pixels have a value of 255 (white on grayscale) or (255,255,255) (white on RGB) and are excluded from altitude evaluation. This 2D altitude image, converted into image form, is an observed input image $\mathbf{X}_j$, $j = 1, ...M$. The corresponding satellite image (depicting actual geomorphology of the patch region) is employed as



(a)

(b)

(c)

(d)

Fig. 1: Four examples of GAN-terrain input/ground-truth/prediction triplets, having NHI similarity scores: a) 0.8393, b) 0.8824, c) 0.9482, d) 0.7944.

ground-truth output image $\mathbf{Y}_j$, $j = 1, ...M$. Thus, the training dataset is constructed by pairs $\{\mathbf{X}_j\ ,\ \mathbf{Y}_j\}$, $j = 1, ...M$.

Each 2D altitude image $\mathbf{X}_j$ can be constructed in two slightly different ways: a) a grayscale one-channel image can be derived by encoding normalized altitude R_{ji} per-PoI as a pixel luminance value. Alternatively, a linear color palette can be used to convert normalized altitude R_{ji} into RGB color values, in order to finally obtain a three-channel colored image (e.g., one from blue to yellow, where the deepest blue/yellow denotes sea level/highest mountain peak level, respectively). Both approaches were implemented and compared in the context of this paper, as described in Section III.

As shown in Figure 1, the visual properties of the generated content are correlated with the color-coded altitude of the input PoIs; in all other respects GAN-terrain has realistically filled-in the generated terrain details fully autonomously. At model deployment-time, random input altitude images can be constructed very rapidly on-the-fly in an automated manner, thanks to the very minimal amount of required information. Even manually drawn, swiftly sketched arbitrary images can be utilized as inputs; a trained GAN-terrain model will successfully interpret them as altitude images, as shown in Figure 4.

In general, output diversity is an important property of a successful PCG system. In the GAN-terrain case, the purpose of the final trained GAN model during system deployment is not to precisely translate the input altitude image into an actual satellite image, but to procedurally generate a new, realistic but imaginary terrain, which may be only vaguely based on the given input. Thus, in order to enhance trained model output diversity, we optionally perform random rotations and/or flipping of each $\mathbf{X}_j$, $j = 1, ...M$ to augment the training dataset, without changing the corresponding $\mathbf{Y}_j$, $j = 1, ...M$. Below, we refer to GAN-terrain models trained with/without this optional augmentation strategy as "Augmented"/"Non-augmented", respectively.

As shown in Section III, this training set augmentation strategy allows the final GAN to synthesize terrain images of greater apparent diversity, by forcing it to ignore input orientation during training. Thus, during deployment of the trained model, small rotations to the input altitude image may produce arbitrarily large rotations to the output, since output orientation is in fact arbitrarily "decided" by the model and not constrained by input orientation. Thus, the Non-augmented model is forced more intensely to mimic ground-truth, while the Augmented one typically provides a more diverse result.

III. GAN-terrain Evaluation

We employed publicly available geographical data [15] in order to construct the training and testing sets for our method. We initially collected $N = 11.2$ million world PoIs, which were utilized to create $M = 4300$ geographic patches and attach their corresponding satellite images (of 512×512 pixels resolution) using the Microsoft Bing Maps API.

The employed GAN architecture was based on the Pix2Pix Network [14]. The network was trained using 3000 input/output patch pairs $\{\mathbf{X}_j\ ,\ \mathbf{Y}_j\}$ and was evaluated using a test set of 1300 input/output patch pairs. Color and grayscale variants of the dataset were used for training separate GAN-terrain models. Color 2D altitude images resulted in predicted network outputs with a higher level of detail than the ones obtained using grayscale inputs, thus GAN-terrain evaluation proceeded with the color variant only. The results were impressive, as GAN-terrain successfully created highly realistic complex terrain images from very simplistic inputs. Training was completed in 300 epochs, on a 24-core Intel Xeon PC with 256GB RAM and an NVIDIA GeForce GTX2080Ti GPU.

Evaluating the quality of synthesized images is an open and difficult problem [16]. In this paper we chose a simple objective evaluation approach, exploiting the fact that pixel color in the output image encodes semantic information. Thus, we measured the Normalized Histogram Intersection similarity (NHI) [17] between the 64-bin joint HSV color histograms [18] of each GAN-terrain prediction and its corresponding ground-truth image from the test set. Minimum/maximum NHI similarity values are $0.0/1.0$, respectively. High NHI similarity between the ground-truth and predicted image histograms can be interpreted as high *semantic concordance* among them, with regard to the distribution of visible geomorphological details (water bodies, forest, snow, mountains, etc.).

Quantitative results indicate that the mean NHI similarity between 1300 ground-truth and predicted images is indeed relatively high (0.7665). This implies that, when the trained GAN model is given a previously unseen 2D altitude image, it synthesizes a highly similar terrain image in terms of semantic concordance. Although NHI similarity does not capture differences between the two terrain images in terms of the the exact landmass/coastline shape/orientation, this is rather irrelevant to terrain image generation task, since our goal is not to replicate the ground-truth terrain. Examples of altitude image, ground-truth terrain image and predicted output image triplets are presented in Figure 1. NHI similarity for each triplet is included for visual inspection purposes.The occasional phenomenon of semantic disconcordance between ground-truth and predicted image (example in Figure 2) can be attributed to the minimal information content of the input altitude images, which is however the main advantage of the proposed method: such 2D scatter plot inputs can be rapidly and easily constructed in model deployment-time.

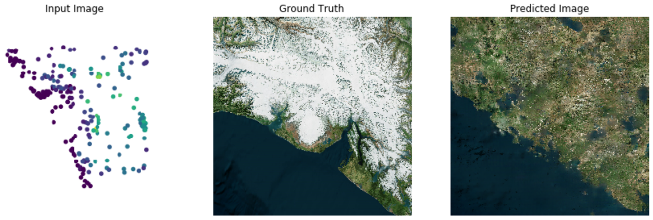


Fig. 2: Test set example of semantic disconcordance between prediction and ground-truth. Here, the network avoided to synthesize snow (NHI score: 0.58). However, the predicted terrain image is still realistic-looking.

Additionally, we performed a subjective evaluation of generated terrain images, using 40 terrain images from our test set and 10 observers. The goal of the subjective evaluation was to let observers deduce in a systematic manner: a) whether the predicted terrain images resembles a real satellite terrain image ("plausibility"), and b) the spatial correspondence between the input 2D altitude image PoIs and the predicted terrain image ("correspondence"). We employed 20 predicted generated terrain images shuffled with 20 ground-truth terrain images for control purposes, totalling 40 terrain images. The participating subjects did not know whether each terrain image they saw was a ground-truth or a predicted one. For each image, they recorded two integer score values in the range $[1, 5]$ for plausibility and correspondence evaluation, respectively.

TABLE I: Evaluation results of the Non-augmented GAN-terrain model. Correspondence and plausibility are scored using a scale in $[1, 5]$, while NHI similarity is a percentage. In all cases higher is better.

Type	**Mean NHI Similarity**	**Mean Correspondence**	**Mean Plausibility**
Predicted Image	0.7665	4.6633	4.4682
Ground Truth	N/A	4.6138	4.5955

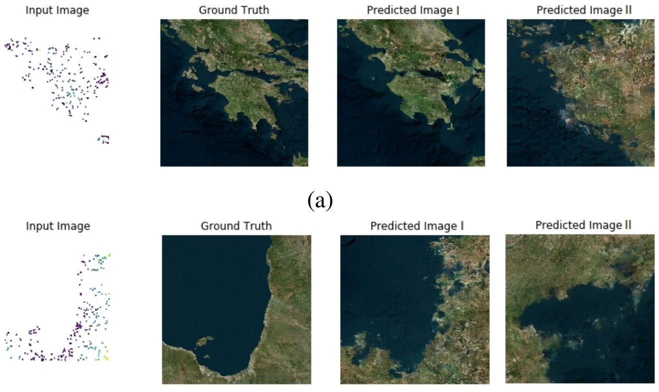
TABLE II: Predicted images diversity comparison between the Non-augmented and Augmented model, using GIST descriptors and total variance.

Measure	**Non-augmented**	**Augmented**
Trace of GIST cov. matrix	0.18995	0.23454
Mean of the main diagonal of GIST cov. matrix	0.000207	0.000244
Variance of the main diagonal of GIST cov. matrix	3.33e-8	3.55e-8
Mean of NHI similarities of joint HSV histograms between ground-truth and predictions	0.7665	0.74172

Subjective evaluation results, shown in Table I, indicate that ground-truth and predicted images are nearly indistinguishable by human subjects: mean correspondence for predicted/ground-truth images was 4.6633/4.6138, respectively, while mean plausibility for predicted/ground-truth images was 4.4682/4.5955, respectively. In fact, artificial GAN-terrain images performed even better than the real ones.

Subjective evaluation was necessarily performed with a GAN model trained using the non-augmented training dataset variant of the proposed method, due to the nature of the employed "correspondence" qualitative metric. Disabling the proposed training data augmentation strategy, which was described in Section II, imposes shape/orientation constraints to be learned by GAN-terrain. Thus, absence of training dataset augmentation may reduce the diversity of GAN-terrain outputs during deployment. To quantify this possibility, we trained a second GAN-terrain model using training data augmentation and then compared the predictions of the two GAN-terrain models on the test set. Evaluation consisted in calculating a GIST global image description vector [19] for each predicted terrain image in the test set, once for the Non-augmented and once for the Augmented model, and subsequently computing the mean global dispersion of these descriptors. This can be measured by averaging over the total variance (i.e., trace of the covariance matrix) of the 1300 960-dimensional GIST vectors $\mathbf{f}_i, i = 1, ..., 1300$, separately for the two models.

The results, shown in Table II, indicate that the mean global dispersion/total variance of test set predictions is significantly greater on the Augmented model variant, where our input augmentation strategy was enabled during training: it is 0.23454/0.18995 for the Augmented/Non-Augmented variant, respectively. To grasp a sense of the significance of this



(a)

(b)

Fig. 3: Outputs of pre-trained Non-augmented/Augmented model (”Predicted Image I/II”, respectively) using two different altitude image inputs from the test set: a) and b).

difference in total variance magnitude, we report that the mean/variance of the main diagonal of GIST covariance matrix in the Augmented variant is 0.000244/3.55e-8, respectively. On the other hand, mean NHI similarity of joint HSV histograms between ground-truth and predictions is slightly higher for the Non-augmented GAN-terrain model: it is 0.7665, versus 0.74172 for the Augmented case. This indicates a slight trade-off between semantic concordance and output diversity.

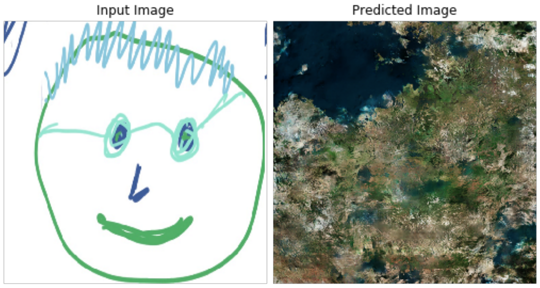


Fig. 4: An example input/predicted image pair, using: a) the pretrained Augmented GAN-terrain model, and b) an arbitrary input (smiling face with glasses), manually drawn in less than 30 seconds.

IV. Conclusions

The proposed GAN-terrain method is able to derive realistic 2D terrain images resembling satellite images in model deployment-time, given only 2D altitude images containing rough PoI scatter plots that encode spatial distribution and altitude of desired geographic landmarks. Although the altitude images employed for training were constructed using real geographic data, similar arbitrary input images can easily and rapidly be created at the inference stage using any image processing software. In contrast, all competing GAN-based terrain generation methods require more sophisticated deployment-time inputs that are comparatively difficult to construct. The output images can be easily transformed into semantically rich 3D terrain meshes by trivial post-processing. Extensive evaluation of the generated terrain images indicates a relatively high degree of semantic concordance between the expected terrain geomorphology and the actually GAN-terrain generated ones, as well as very realistic and plausible generated terrains. Additionally, GAN-terrain evaluation results indicate a predicted terrain image diversity gain, at a very low penalty in semantic concordance, when using the proposed training data augmentation strategy. Future work may focus on actually synthesizing 3D terrain content, generating both terrain image texture and geometry data.

References

[1] S. Worley, “A cellular texture basis function,” in *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques*, 1996.

[2] S. Gustavson, “Simplex noise demystified,” *Linköping University, Linköping, Sweden, Research Report*, 2005.

[3] T. Archer, “Procedurally generating terrain,” in *Proceedings of the Annual Midwest Instruction and Computing Symposium*, 2011.

[4] G.S.P. Miller, “The definition and rendering of terrain maps,” in *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques*, 1986.

[5] J. Doran and I. Parberry, “Controlled procedural terrain generation using software agents,” *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 2, no. 2, pp. 111–119, 2010.

[6] J. Togelius, G. N. Yannakakis, K. O. Stanley, and C. Browne, “Search-based procedural content generation: A taxonomy and survey,” *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 3, no. 3, pp. 172–186, 2011.

[7] X. Mei, P. Decaudin, and B.-G. Hu, “Fast hydraulic erosion simulation and visualization on GPU,” in *Proceedings of the Pacific Conference on Computer Graphics and Applications*. IEEE, 2007.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative Adversarial Nets,” in *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 2014.

[9] J. Klein, S. Hartmann, M. Weinmann, and D. L. Michels, “Multi-scale terrain texturing using Generative Adversarial Networks,” in *Proceedings of the International Conference on Image and Vision Computing New Zealand (IVCNZ)*. IEEE, 2017.

[10] R. J. Spick, P. Cowling, and J. A. Walker, “Procedural generation using spatial GANs for region-specific learning of elevation data,” in *Proceedings of the IEEE Conference on Games (CoG)*, 2019.

[11] C. Beckham and C. Pal, “A step towards procedural terrain generation with GANs,” *arXiv preprint arXiv:1707.03383*, 2017.

[12] R. Spick and J. Walker, “Realistic and textured terrain generation using GANs,” in *Proceedings of the European Conference on Visual Media Production (CVMP)*, 2019.

[13] E. Panagiotou and E. Charou, “Procedural 3D terrain generation using Generative Adversarial Networks,” *arXiv preprint arXiv:2010.06411*, 2020.

[14] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[15] www.geonames.org, “https://www.geonames.org,” 2018.

[16] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved techniques for training gans,” in *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 2016.

[17] E. Cheng, N. Xie, H. Ling, P. R. Bakic, A.D.A. Maidment, and V. Megalooikonomou, “Mammographic image classification using histogram intersection,” in *Proceedings of the IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2010.

[18] S. Sural, A. Vadivel, and A. K. Majumdar, “Histogram generation from the HSV color space,” in *Encyclopedia of Information Science and Technology, First Edition*, pp. 1333–1337. IGI Global, 2005.

[19] M. Douze, H. Jégou, H. Sandhawalia, L. Amsaleg, and C. Schmid, “Evaluation of GIST descriptors for web-scale image search,” in *Proceedings of the ACM International Conference on Image and Video Retrieval*, 2009.