# SeamlessGAN: Self-Supervised Synthesis of Tileable Texture Maps

Carlos Rodriguez-Pardo, Elena Garces

**Abstract**—Real-time graphics applications require high-quality textured materials to convey realism in virtual environments. Generating these textures is challenging as they need to be visually realistic, seamlessly tileable, and have a small impact on the memory consumption of the application. For this reason, they are often created manually by skilled artists. In this work, we present *SeamlessGAN*, a method capable of automatically generating tileable texture maps from a single input exemplar. In contrast to most existing methods, focused solely on solving the synthesis problem, our work tackles both problems, *synthesis* and *tileability*, simultaneously. Our key idea is to realize that tiling a latent space within a generative network trained using adversarial expansion techniques produces outputs with continuity at the seam intersection that can then be turned into tileable images by cropping the central area. Since not every value of the latent space is valid to produce high-quality outputs, we leverage the discriminator as a perceptual error metric capable of identifying artifact-free textures during a sampling process. Further, in contrast to previous work on deep texture synthesis, our model is designed and optimized to work with multi-layered texture representations, enabling textures composed of multiple maps such as albedo, normals, etc. We extensively test our design choices for the network architecture, loss function, and sampling parameters. We show qualitatively and quantitatively that our approach outperforms previous methods and works for textures of different types.

**Index Terms**—Artificial intelligence, Artificial neural network, Machine vision, Image texture, Graphics, Computational photography

## 1 Introduction

Realistic and high-quality textures are important elements to convey realism in virtual environments. These can be procedurally generated [1], [2], [3], [4], [5], [6], [7], captured [8], [9] or synthesized from real images [10], [11], [12], [13]. Frequently, textures are used to efficiently reproduce elements with repetitive patterns (for example, facades, surfaces, or materials) by means of spatially concatenating –or *tiling*– multiple copies of themselves. Creating tileable textures is a very challenging problem, as it requires a semantic understanding of the repetitive elements, often at multiple scales. For this reason, such a process is frequently done manually by artists in 3D digitization pipelines.

Recent advances in Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) have been applied to texture synthesis problems [12], [14], [15], [16], [17], [18], [19] showing unprecedented levels of realism and quality, however, the output of these methods is not tileable. Despite recent methods [13], [16], [20], [21], [22], [23] addressing the problem of tileable texture synthesis, we show that they either assume a particular level of regularity or the generated textures lose a significant amount of visual fidelity with respect to the input exemplars. Further, most of these methods have only focused on synthesizing single images. Rendering realistic materials requires more information about their optical properties beyond what represents a single RGB pixel. To this end, it is common to use spatially-varying BRDFs [24], which are optical appearance models parameterized by stacks of images, each one representing a different property, such as albedo, normals, or transparency. As the number of methods that generate texture stacks from physical samples grows [24], [25], [26] so does the need to turn them into *tileable texture stacks*.

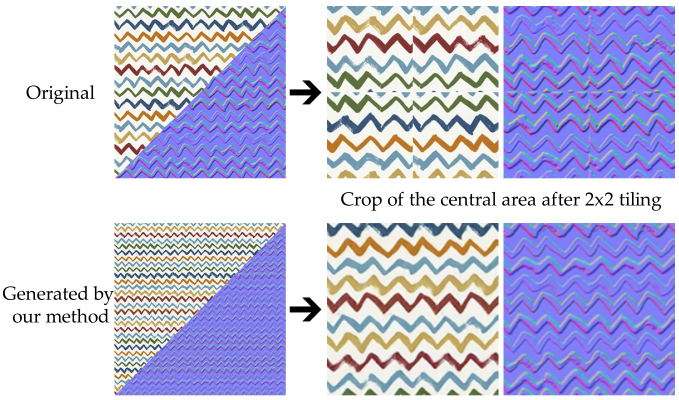



Fig. 1: Naïvely tiling the original texture causes discontinuities at the seam intersections as shown in the top row. Our method automatically generates a tileable texture stack from an input exemplar which double the size of the input.

In this work, we propose a deep generative model, *SeamlessGAN*, capable of synthesizing tileable texture stacks from inputs of arbitrary content. In contrast to *Wang Tiles* [27], by which a single large texture region is created by concatenating multiple different tiles with matching borders, we aim to automatically obtain a seamless single-tile, which allows for reduced memory consumption and enhanced usability in casual scenarios when the user might lack the necessary artistic skills. Our key idea is to realize that tiling a latent space within a generative network produces outputs with continuity at the seam intersection [14], which can then be turned into tileable images by cropping the central area. Since not every value of the latent space is valid to produce high-quality outputs, we follow a double strategy: First, we train the generative network using an adversarial expansion technique [12], which provides a latent encoding of the input texture, which can then be decoded into high-quality outputs that

- *Carlos Rodriguez - Pardo is with SEDDI (28007, Madrid, Spain) and with Universidad Carlos III de Madrid (28005, Madrid, Spain). E-mail: carlos.rodriguezpardo.jimenez@gmail.com*
- *Elena Garces is with SEDDI (28007, Madrid, Spain) and with Universidad Rey Juan Carlos (28933, Madrid, Spain) E-mail: elena.garces@seddi.com*

double the spatial extent of the input. Second, we use the trained discriminator as a local quality metric in a sampling algorithm. This allows us to find the input of the generative process that produces tileable textures similar to the original, as well as multiple candidates per input exemplar. As opposed to previous work, which focused on maximizing stationarity [13] and thus might remove important high-level texture features in regular or near-regular textures, our method is focused on maximizing tileability while preserving the original texture as intact as possible, in terms of its stylistic and semantic properties. To allow for the synthesis of stacks of textures, we propose a neural architecture composed of various decoder networks. Despite not explicitly imposing inter-map consistency, we show that it is implicitly guaranteed by how the generative network is trained. Without losing generality, we show texture stacks synthesis of two maps: an albedo and a normal map. We demonstrate that our method outperforms state-of-the-art solutions on tileable texture synthesis of single images and show several examples for synthesizing tileable texture stacks. We validate our design choices through several ablations studies and off-the-shelf perceptual quality metrics.

## 2 Related work

We review the texture synthesis methods most closely related to our work. For a more comprehensive review, please refer to the surveys in [28], [29]. We will also mention other related work regarding tileable texture synthesis, and deep internal learning.

### 2.1 Texture Synthesis

Traditionally, **non-parametric texture synthesis** algorithms worked by ensuring that every patch in the output textures approximates a patch in the input texture. Earlier methods included image quilting [10], [30], GraphCuts [31], genetic algorithms [32], and optimization [11], [33]. More recent approaches use variations of PatchMatch [34], [35] as a way of finding correspondences between generated and input images [36], [37], [38].

Despite those methods showing high-quality results for textures of different characteristics, recent work on deep parametric texture synthesis show better generality and scalability, requiring less manual input. Our approach belongs to the category of **Parametric texture synthesis**. These methods learn statistics from the example textures, which can then be used for the generation of new images that match those statistics. While traditional methods used hand-crafted features [39], [40], recent parametric methods rely on deep neural networks as their parameterization. Activations within latent spaces in pre-trained CNNs have shown to capture relevant statistics of the style and texture of images [41], [42], [43], [44], [45]. Textures can be synthesized through this approach by gradient-descent optimization [46], [47] or by training a neural network that learns those features [23], [48]. Finding generic patterns that precisely describe the example textures is one of the main challenges in parametric texture synthesis. Features that describe textured images in a generic way are hard to find and they typically require hand-tuning. Generative Adversarial Networks (GANs), which have shown remarkable capabilities in image generation in multiple domains [49], [50], [51], can learn those features from data. Specifically, in texture synthesis, they have proven successful at generating new samples of textures from a single input image [12] or from a dataset of images [14], [15], [16], [17], [18]. We build upon the method of Zhou *et al.* [12] which shows good performance on the synthesis of non-stationary single image textures, and extend it to synthesize texture stacks, as well as generate tileable outputs.

### 2.2 Texture Tileability

Synthesizing tileable textures has received surprisingly little attention in the literature until recent years. Moritz *et al.* [13] propose a non-parametric approach that is able to synthesize textures from a single example while preserving its *stationarity*, which measures how tileable the texture is. Li *et al.* [22] propose a GraphCuts-based algorithm. They first find a patch that optimally represents the texture, then use graph cuts to transform its borders to improve its tileability. This method allows for synthesizing multiple maps at the same time. Relatedly, Deliot and Heitz [20] propose a histogram-preserving blending operation for patch-based synthesis of tileable textures, particularly suited for stochastic textures. The power of deep neural networks for tileable texture synthesis has also been leveraged in the past years. First, Rodriguez-Pardo *et al.* [21] exploit latent spaces in a pre-trained neural network to find the size of the repeating pattern in the input texture. Then, they use perceptual losses for finding the optimal crop of the image such that, when tiled, looks the most similar to the original image. Also leveraging deep perceptual losses and using Neural Cellular Automata [52] as an image parameterisation, Niklasson *et al.* [23] generate self-organizing textures that are seamlessly tileable by design, but are limited in resolution and by the quality of the gram matrix perceptual metric used as a loss function [53]. Bergmann *et al.* [16] achieve tileability in their output textures by training a multi-image GAN in a periodic spatial manifold. Our proposed method does not follow any of these approaches. Instead, we build upon a state-of-the-art single-image texture synthesis method, which is able to generate high-quality and high-resolution images, and extend it to generate textures which are tileable. To this end, we propose a sampling algorithm that finds the input of the generative process that maximizes a novel tileability metric. Our goal is to preserve the input original texture as intact as possible while imposing artifact-free continuity at the intersection of the seams when the texture is tiled. Furthermore, our method has a reduced computational footprint compared to other deep generative texture synthesis methods as we say in the results section.

### 2.3 Deep Internal Learning

Learning patterns from a single image has been studied in recent years, in contexts different to those of texture synthesis. Ulyanov *et al.* [54] show that single images can be represented by randomly initialized CNNs, and show applicability on denoising or image inpainting problems. A similar method is proposed by Shocher *et al.* [55] for single-image super-resolution. Additionally, single images have shown to be enough for learning low-level features that generalize to multiple problems [56]. Hypernetworks [57] in *Implicit Neural Representations* [58], [59] also allow for shift-invariant priors over images, which can then used in generative models [60]. Single-image generative models have been explored for domains other than textures. GANs trained on a single image have been used for image retargeting [61], deep image analogies [62], or for learning a single-sample image generative model [63], [64], [65], [66], [67]. These methods, while powerful for natural images, are not well-behaved for textures, as shown in [15] and [18]. By introducing inductive biases specially designed for textured images, characterized by their repeating patterns, deep texture synthesis methods achieve better performance in textured images than generic single-image synthesis approaches, which need to account for more globally coherent semantic patterns.

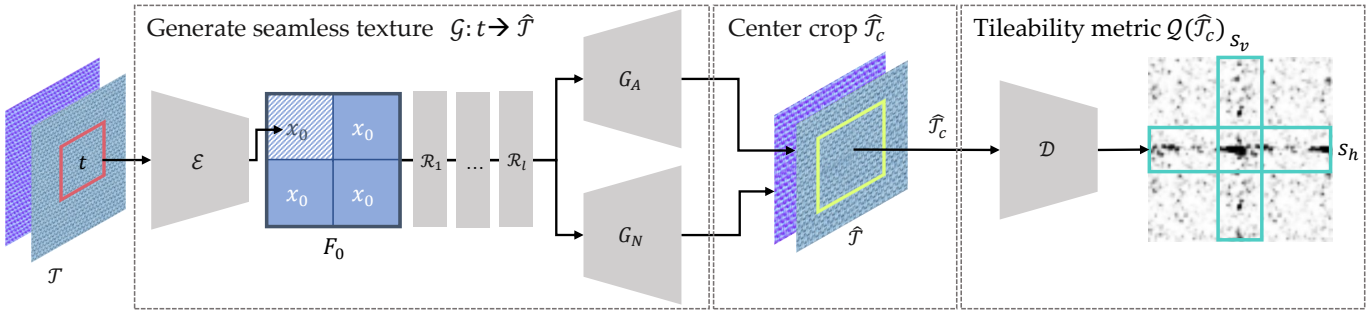



Fig. 2: Overview of *SeamlessGAN*. A crop $t$ of the input texture stack $\mathcal{T}$, is fed to an encoder $\mathcal{E}$, which transforms it into a latent space $x_0$. We tile this latent space vertically and horizontally, obtaining a latent field $\mathrm{F}_0$. $\mathrm{F}_0$ is further processed by several residual blocks $\mathcal{R}_i, i \in \{1, l\}$. The resulting latent variables are transformed by two different decoders, $\mathrm{G}_A$ and $\mathrm{G}_N$, which output 4 copies of a candidate tileable texture $\hat{\mathcal{T}}$. By cropping the center part of this texture, we obtain a single of those copies, with seamless borders, $\hat{\mathcal{T}}_c$. Additionally, this texture can be analyzed by a discriminator $\mathcal{D}$ which provides local estimations of the quality of the synthesis. Using a tileability evaluation function $\mathcal{Q}$, which, by analyzing two vertical and horizontal search areas $s_v, s_h$, it is able to detect artifacts that may arise when tiling the texture. This gives us an estimation of how *tileable* the texture is. This estimation can then be used by a sampling algorithm for generating high-quality tileable textures.

## 3 Overview

Our method takes as input an *untileable texture stack* $\mathcal{T}$, which is a layered representation –or SVBRDF [68]– of a material used by render engines to virtually reproduce it. A typical texture stack is composed of several maps such as albedo, normals, or roughness. For simplicity, and without losing generality, we assume that our texture stacks have two maps: an albedo A, and a normal map N, both are RGB images of the same dimensions. There are several methods that generate texture stacks for any material using, for example, smartphones [24], [25], [26], [69], however, these stacks are not tileable by default. Having this data as input, we propose a two-step automatic pipeline to generate tileable texture stacks from arbitrary material input exemplars.

In the first step, described in Section 4, we use crops $t$ of the input texture $\mathcal{T}$ to train a Generative Adversarial Network (GAN) able to synthesize novel *untileable* texture stacks $\mathcal{T}'$ using adversarial expansion [12]. This training framework has shown to provide state-of-the-art results on single-sample texture synthesis, surpassing previous approaches [16], [70], [71]. Thanks to using a GAN, we learn an implicit representation of the texture parameterised in two neural modules: $\mathcal{G}$ and $\mathcal{D}$. $\mathcal{G} : t \rightarrow \mathcal{T}'$ is a generator that outputs new untileable textures which double the spatial resolution of the input. $\mathcal{D}$ is a discriminator that, thanks to the adversarial framework used for training, is able to distinguish real from fake textures.

In the second step, described in Section 5, we produce tileable stacks by means of two key ideas: first, we tile a latent space $\mathrm{x}_0$ of the trained generator $\mathcal{G}$, obtaining a novel texture stack showing continuity at the seams intersection $\hat{\mathcal{T}}$. Second, we implement a sampling process using the trained discriminator $\mathcal{D}$ used as quality metric $\mathcal{Q}$ to find an optimally tileable texture stack $\mathcal{T}^*$. An overview of our sampling step is shown in Figure 2.

In the following, we first describe our GAN architecture, including our proposal to deal with textures stacks. Then, we describe the sampling process to generate tileable ones.

## 4 Self-Supervised Texture Stack Synthesis using Adversarial Expansion

For each input texture $\mathcal{T}$ we train a GAN, whose generator $\mathcal{G}$ is able to synthesize novel examples $\mathcal{T}'$. Unlike other GAN frameworks, which take as input a random vector, we use crops $t$ of the original stack as input to guide the generative sampling, such that, $\hat{\mathcal{T}}' = \mathcal{G}(t)$ for $t \in \mathcal{T}$. The training strategy builds upon the work of Zhou *et al.* [12], which uses *adversarial expansion* to train the network as follows: First, a target crop $t \in \mathcal{T}$ of $2k \times 2k$ pixels is selected from the input stack. Then, from that target crop $t$, a source random crop $t_s \in t$ is chosen with a resolution of $k \times k$ pixels. The goal of the generative network will be to synthesize $t$ given $t_s$. This learning approach is fully self-supervised. The generative model is trained alongside a discriminator $\mathcal{D}$, which learns to predict whether its inputs are the target crops $t \in \mathcal{T}$ or the generated samples $\mathcal{T}' = \mathcal{G}(t_s)$. Figure 3 shows an overview of the training strategy.

### 4.1 Network Architecture

SeamlessGAN is comprised by an encoder-decoder convolutional generator $\mathcal{G}$ with residual connections, and a convolutional discriminator $\mathcal{D}$. So as to be able to synthesize textures of multiple different sizes, the networks are designed to be fully-convolutional. We follow the residual architectural design in [12], with two extensions: First, building on recent advances on style transfer algorithms, we use *Instance Normalization* [72] before each ReLU non-linearity in the generator and each Leaky-ReLU [73] operation in the discriminator. This allows to use normalization for training the networks without the typical artifacts caused by Batch Normalization [74], [75], [76], [77]. Second, in order to allow for the synthesis of multiple texture maps at the same time, we propose a variation in the generator architecture. In the following, we describe the details of each component, with a particular focus on the elements that are different from [12].

**Generator**: The generative architecture proposed is comprised of three main components: An *encoder* $\mathcal{E}$, which compresses the information of the input texture $t$ into a latent space, $\mathrm{x}_0 \leftarrow \mathcal{E}(t)$, with half the spatial resolution of the input texture. Then, a set of *residual blocks*, $\mathcal{R}_i, i \in \{1, l\}$, which learn a compact representation of the input texture $\mathrm{x}_i \leftarrow \mathcal{R}_i(\mathrm{x}_{i-1}) + \mathrm{x}_{i-1}$. Finally, a stack of *decoders* $\mathbf{G}$ that transforms the output of the last residual block $\mathcal{R}_l$ into an output texture $\mathcal{T}' \leftarrow \mathbf{G}(\mathrm{x}_l)$. Residual learning allows for training deeper models with higher levels of visual abstraction and which generate sharper images [78], [79], [80]. Similar residual generators have been used in unsupervised image-to-image translation problems [81].

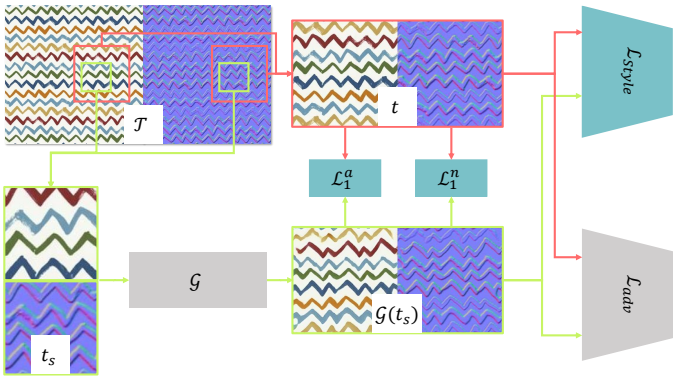


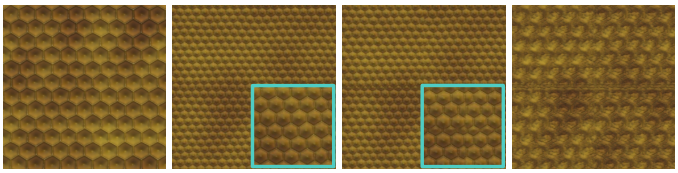

Fig. 3: An overview of our training framework for learning to synthesize texture stacks through adversarial expansion. At each iteration, from an input stack $\mathcal{T}$, we randomly crop a target crop $t$, from which we select a source crop $t_s$. The goal of the generator $\mathcal{G}$ is to estimate $t$ given $t_s$. We measure the difference between target and estimated crops using a combination of adversarial, pixel-wise and perceptual loss functions.

Synthesizing a texture stack of multiple maps with a single generative model poses extra challenges over the single map case. Each map represents different properties of the surface, such as geometry or color, resulting in visually and statistically different images. This suggests that independent generative models per map could be needed. However, the texture maps must share pixel-wise coherence, which is not achievable if multiple generative models are used. Inspired by previous work on intrinsic images [82], [83], [84], we propose the use of a generative model that learns a shared representation of all the texture maps, but has different decoders for each of them. The latent space within the generator will thus encode a high-level representation of the texture, which is then decoded in different ways for each different map in the texture stack. Specifically, we train a stack of decoders, $\mathbf{G} = \{G_{\mathrm{A}}, G_{\mathrm{N}}\}$, one for each map of the stack (albedo and normals in our case).

**Discriminator**: For the discriminator $\mathcal{D}$, we use a *PatchGAN* architecture [12], [80], [81], [85], which, instead of providing a single estimation of the probability of the whole image being real, it classifies this probability for small patches of it. This architecture has several advantages for our problem. First, it provides local assessments on the quality of the synthesized textures, which we exploit for obtaining high-quality textures. Second, its architecture design allows to provide some control on what kind of features are learned: by adding more layers to $\mathcal{D}$, the generated textures are typically of a higher semantic quality but local details may be lost, as we show on the Supplementary Material. A comprehensive study on the impact of the depth of $\mathcal{D}$ can be found in [12].

**Loss Function**: We train the networks following a standard GAN framework [86]. We iterate between training $\mathcal{D}$ with a real sample texture stack $\mathcal{T}$ and a generated sample $\mathcal{T}'$. The adversarial loss $\mathcal{L}_{adv}$ is extended with three extra loss terms: $\mathcal{L}_1^a$, $\mathcal{L}_1^n$, and $\mathcal{L}_{style}$; corresponding respectively to the pixel-wise distance between the generated and target albedo, normals, and a perceptual loss. The perceptual loss $\mathcal{L}_{style}$ is computed as the sum of the perceptual losses between target and generated albedo maps, and target and generated normal maps. We follow the gram loss described in [87] as our perceptual loss. We weight the total style loss by weighting different layers in the same way described in [12], [21], [87]. Our global loss function thus is defined as:

$$\mathcal{L} = \lambda_{adv}\mathcal{L}_{adv} + \lambda_{\mathcal{L}_1^a}\mathcal{L}_1^a + \lambda_{\mathcal{L}_1^n}\mathcal{L}_1^n + \lambda_{style}\mathcal{L}_{style} \quad (1)$$



(a) $t$ (b) $\hat{\mathcal{T}}, l = 0$ (c) $\hat{\mathcal{T}}, l = 3$ (d) $\hat{\mathcal{T}}, l = 5$

Fig. 4: Impact of the latent field level $l$ on the quality of the output textures $\hat{\mathcal{T}} = \mathcal{G}(t)$, using the same input $t$. Generating the latent field $\mathrm{F}_l$ early layers ($l = 0$) generates the best visual results, later layers either create small artifacts ($l = 3$) or generate unrealistic textures ($l = 5$). A zoom of the central crop is shown as an inset.

## 5 Tileable Texture Stack Sampling

### 5.1 Latent Space Tiling

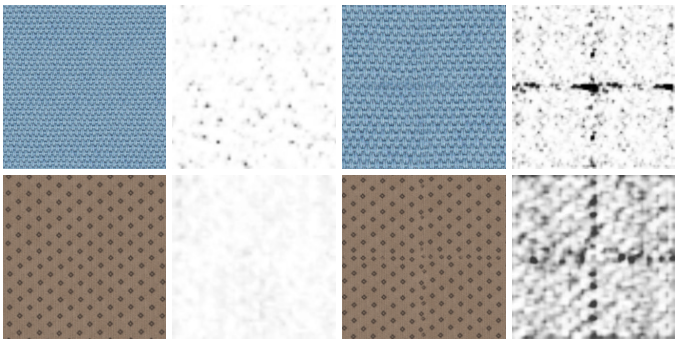
After training, the generator $\mathcal{G}$ is able to synthesize novel samples of the texture given a small exemplar of it. Although these novel samples duplicate the size of the input, they are not tileable by default. Previous work [14] showed that by spatially concatenating different latent spaces in a ProGAN [49] generator, it is possible to generate textures that contain the visual information of those tiles while seamlessly transitioning between the generated tiles. Inspired by this idea, we spatially repeat (horizontally and vertically) the first latent space $\mathrm{x}_0$ within the generative model, obtaining a latent field $\mathrm{F}_0$. This field is passed through the residual layers $\mathcal{R}_l$ and the decoders $\mathbf{G}$ to get a texture stack, $\hat{\mathcal{T}}$, that contains four copies of the same texture with seamless transitions between them (see Figure 2). At first, the latent field $\mathrm{F}_0$ shows strong discontinuities at the borders between the four copies. Later, as the texture is passed through the network, these artifacts are progressively transformed into seamless borders by the rest of the residual blocks and the decoder of the model. The intuition behind this idea is that, after training the network with the target texture, each of these latent spaces encode low resolution versions of the input in the spatial domain and semantically-rich information in the deeper layers.

A seamlessly tileable texture stack $\hat{\mathcal{T}}_c$ is obtained by cropping the central region (with an area of $50\%$ of $\hat{\mathcal{T}}$). The predicted texture $\hat{\mathcal{T}}$ has $4 \times 4$ the resolution of the input crop $t$. Thus, after the cropping operation, $\hat{\mathcal{T}}_c$ has twice the resolution of the input $t$.

A key parameter to select is the level $l$ at which to split the generative process. As shown in Figure 4, generating the latent field F by tiling earlier levels of the latent space forces the network to transform F more times, thus resulting in more seamlessly tileable textures. We confirm the results found in [14] and noticed that generating this latent field at earlier levels of the latent space ($l \in \{0, 1\}$) yields the best visual results, by allowing for smoother and more semantically coherent transitions between tiles. We thus tile the output of the encoder $\mathrm{x}_0 = \mathcal{E}(t)$, before transforming it by the residual blocks $\mathcal{R}_i$.

### 5.2 Discriminator-guided Sampling

Our strategy to tile the latent space guarantees that the generated texture is a continuous function with smooth transitions between the boundaries of the tiles. However, in contrast to the algorithm in [14], where the latent spaces are drawn from random vectors, ours are encoded representations of input textures. Thus, the selection of the input that the network receives plays an important role on the quality of the output textures. As shown in Figure 5, not all the generated textures are equally valid. This selection can be posed



(a) $\hat{\mathcal{T}}_1$ (b) $\mathcal{D}(\hat{\mathcal{T}}_1)$ (c) $\hat{\mathcal{T}}_2$ (d) $\mathcal{D}(\hat{\mathcal{T}}_2)$

Fig. 5: Outputs of $\mathcal{D}$ with textures of different qualities. (a) The generated albedos $\hat{\mathcal{T}}_1$ are artifact-free on the search areas (b), thus the discriminator $\mathcal{D}(\hat{\mathcal{T}}_1)$ is fooled to believe $\hat{\mathcal{T}}_1$ are real. (c) Failed generated samples $\hat{\mathcal{T}}_2$ cannot fool the discriminator, which finds artifacts $\mathcal{D}(\hat{\mathcal{T}}_2)$ (d). Further examples can be found on the Supplementary Material.

as an optimization problem: $t^* = \text{argmax}_t \mathcal{Q}(\mathcal{G}(t))$, where the goal is to find the crop $t^*$ that maximizes the quality $\mathcal{Q}$ of the generated texture $\mathcal{G}(t) = \hat{\mathcal{T}}_c$. To solve this optimization problem, one option could be to pose it as an optimization in the generative latent space. However, we found that a simpler solution based on sampling already provides satisfactory results. The remainder of this section explains the sampling process and the quality metric.

**Sampling**: By using a fully-convolutional GAN, our model can generate textures of any sizes, hardware being the only limiting factor. This is key for tileable texture synthesis as, even if a given texture is seamlessly tileable, larger textures require fewer repetitions to cover the same spatial area, which ultimately results in fewer repeating artifacts, as illustrated in Figure 6 (a). There are two main challenges when finding tileable textures: the input texture needs to contain the distribution of the existing repeating patterns; and the input tile itself must not create strong artifacts when tiling the latent spaces of the generator.

Our goal is thus to find the largest possible tileable texture stack. To do so, we sample multiple candidate crops for a given crop size $c \in \{c_{min}, c_{max}\}$, where $c_{max}$ is the resolution of the input $\mathcal{T}$. We sample crop sizes starting at the largest possible size $c = c_{max}$, and stop when we find a suitable candidate according to the tileability metric $\mathcal{Q}$. As shown in Figure 6 (b), this sampling mechanism also allow us to generate multiple tileable candidates for a single exemplar if we choose different parts of $\mathcal{T}$ as input.

**Discriminator-guided Quality Function, $\mathcal{Q}$**: The second component of our sampling strategy is the quality function used to determine whether a stack is tileable or not. We observed that the artifacts appear on vertical and horizontal frames around the center of the textures (Figure 5). This is likely caused by strong discontinuities or gradients on the same areas of the tiled latent spaces, which the rest of the generative network fails to transform into realistic textures. Following recent work on generative models [88], we use the discriminator $\mathcal{D}$ as a semantic-aware error metric that can be exploited for detecting local artifacts in the generated textures. This can be done in our case because the global loss function contains pixel-wise, *style*, and adversarial losses. The adversarial loss learns the semantics of the texture, whereas the $\mathcal{L}_1$ and style losses model color distances or repeated patterns [21]. This combination of loss functions allows to balance textural, semantic and perceptual color properties.

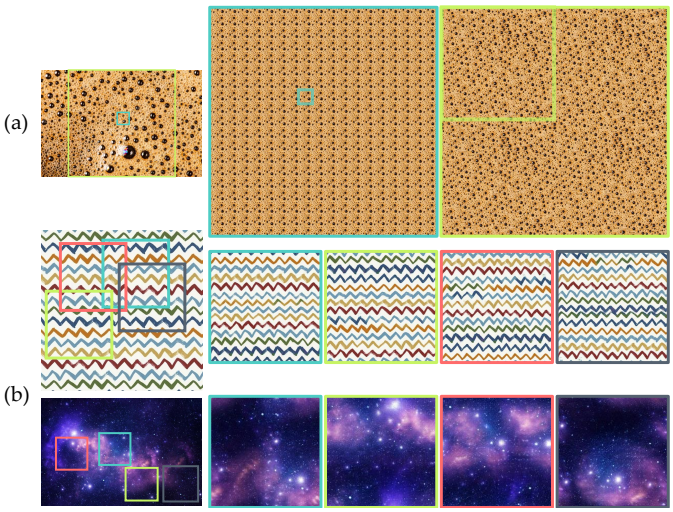



Fig. 6: SeamlessGAN can generate multiple tileable outputs from the same sample $\mathcal{T}$. By cropping different parts of $\mathcal{T}$, we can feed the generator $\mathcal{G}$ with different inputs $t_c$ (shown on the left), generating tileable textures $\hat{\mathcal{T}}_c$ of different sizes. On the example at the top (a), the synthesized images are tiled so they cover the same spatial extent, which shows that, even if the textures are tileable, larger textures generate fewer repeating artifacts. (b) Shows the tile resulting from sampling different parts of the input image. More results are included in the Supplementary Material.

We thus design a quality evaluation function $\mathcal{Q}$ that estimates if the generated texture stack $\hat{\mathcal{T}}_c$ is *tileable*, looking for artifacts on a central area, $S \in \mathcal{D}(\hat{\mathcal{T}}_c)$ of the discriminator. This area is composed of two regions, $S = s_v \cup s_h$, where $s_v$ is a vertical area, and $s_h$ is an horizontal one, both centered on the output of the discriminator (Figure 2). The function $\mathcal{Q}$ leverages the fact that $\mathcal{D}$ outputs 0 when it believes a patch to be synthetic. However, as the values are sample-dependent, we establish a threshold $\tau \in \mathbb{R}$ using the values of the rest of the image $\overline{S}$ as a reference $\tau = \gamma \cdot \min(S_r)$, where $S_r = \mathcal{D}(\hat{\mathcal{T}}_c) \cap \overline{S}$ is the remaining part of the image, and $\gamma \in \mathbb{R}$ is a threshold that allows to control the sensitiveness of $\mathcal{Q}$. Consequently, $\mathcal{Q}(\hat{\mathcal{T}}_c)$ is 1 if $\min(s_v)$ and $\min(s_h)$ are greater or equal than $\tau$, considering the texture as *tileable*, and 0 otherwise. The goal of $\mathcal{Q}$ is to estimate whether or not the central areas, where discontinuities may be present, are as realistic as the rest of the texture. It works as a classification function which returns a high value if the texture is identified as real by the discriminator in such areas. This allows us to create a sampling strategy that distinguishes between images with local artifacts from seamless textures. By using minimum values instead of the mean estimation of the discriminator, our quality metric focuses on detecting artifacts which may arise during the latent space tiling operation.

## 6 Implementation Details

As described in Section 4, we follow a standard GAN training framework, by iterating between training the discriminator and the generator. We use a batch size of 1, and an input size of $k = 128$. All weights are initialized by sampling a gaussian distribution $\mathcal{N}(0, 0.02)$, following standard practice [81]. $\mathcal{G}$ has $l = 5$ residual blocks with ReLU activations, and $\mathcal{D}$ is comprised of 5 convolutional layers with Leaky-ReLU non-linearities, and a Sigmoid operation at the end. We refer the reader to the Supplementary Material for an ablation study on the influence

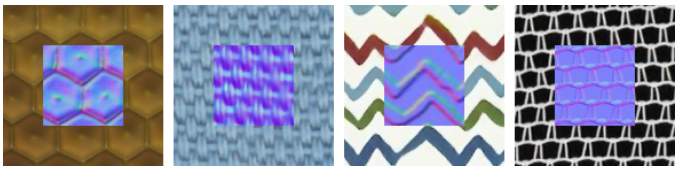


Fig. 7: Crops of synthesized textures using our method. We observe that pixel-wise coherence between maps is preserved.

of the network size. We use a stride of 2 for the downsampling operations in $\mathcal{E}$ and transposed convolutions [89] for upsampling in the decoders $\mathbf{G}$. We weight each part of the loss function as: $\lambda_{adv} = \lambda_{style} = 1$, and $\lambda_{\mathcal{L}_1^a} = \lambda_{\mathcal{L}_1^n} = 10$.

The networks are trained for $50000$ iterations using Adam [90], with an initial learning rate of $0.0002$, which is divided by 5, after iterations $30000$ and $40000$. Aside from random cropping, we do not use any other type of data augmentation method. The models are trained and evaluated using a single NVIDIA GeForce GTX 1080 Ti. Even if training takes around $40$ minutes for each texture stack, once trained, the generator can generate individual new samples in milliseconds, before checking for tileability. We use PyTorch [91] as our learning framework. To accelerate the training process, we leverage mixed precision training and automatic gradient scaling [92]. Every operation in the training pipeline is done natively in GPU using Torchvision [93]. These optimizations allow us to train the networks one order of magnitude faster than previous methods [12]. The input textures $\mathcal{T}$ tested in this paper have, on average, 500 pixels in their larger dimension, for which our method can generate tiles of at most 1000 pixels. We use Photometric Stereo [94] for computing the normals of fabric textures, and artist-generated normals for the other samples. Please refer to the Supplementary Material for further implementation details.

We tile the latent space at its earliest level ($\mathrm{F}_l, l = 0$), as it provides the best quality results, as shown on Section 5.1. For identifying the tileability of textures, we use a search area $S$ that spans a $20\%$ of each spatial dimension of the textures. For all the results shown, $\gamma = 1$. We choose $c_{min} = 100$ pixels, and $c_{max}$ equal to the resolution of the whole input texture. This means that for the level of $c_{max}$, only one texture is sampled. For each $c < c_{max}$, we sample 3 different random crops. Typically, for the results shown in the paper, a tileable texture stack is found at high-resolution crop sizes, $c \geq c_{max} - 10$, thus needing to sample and evaluating less than thirty random crops before finding a satisfactory solution. This entire sampling procedure takes less than five minutes.

## 7 Experiments

In this section, we evaluate our two main contributions: first, in Section 7.1 we study the impact of the design choices of the generator $\mathcal{G}$ to synthesize a multi-layer texture stack, and second, the quality of the tileable texture synthesis through several ablation studies. We study the inter-map consistency in Section 7.2, and the impact of the loss function in Section 7.3. Finally, we compare our method with methods on tileable texture synthesis (Section 8).

### 7.1 Generator $\mathcal{G}$ Design

One of the main challenges when synthesizing texture stacks using a single generator is to preserve the low-level details of each of the texture maps whilst maintaining the local spatial coherence between them; if this coherence is lost, renders that use the synthesized maps will show artifacts or look unrealistic. We propose two different variations of the single-map generative architecture $\mathcal{G}$ presented in [12], each of which makes different assumptions on how the synthesis should be learned taking into account the particular semantics and purpose of each map. A diagram of each proposed architecture is shown in Figure 9. For a fair evaluation, we follow the criteria that both networks must have approximately the same number of trainable parameters.

| | | SSIM ↑ | Si-FID ↓ | LPIPS ↓ |
|---|---|---|---|---|
| $\mathcal{G}_1$ | Albedo | 0.2774 | 0.3462 | 0.4826 |
| | Normals | 0.2364 | **0.3636** | 0.4870 |
| $\mathcal{G}_2$ | Albedo | **0.3123** | **0.3275** | **0.4377** |
| | Normals | **0.2921** | 0.4019 | **0.4340** |

TABLE 1: Quantitative comparison between our variations of the generator $\mathcal{G}$. We show the average results on different metrics across different texture stacks, separated by maps. As shown, $\mathcal{G}_2$ outperforms its baseline across metrics and texture maps. $\mathcal{G}_1$ only yields better scores at the Si-FID metric on the normal map. Higher is better for SSIM, while lower is better for Si-FID and LPIPS.

Our baseline, $\mathcal{G}_1$, treats the texture stack as a multiple channel input image, and entangles every texture map in the same layers. It assumes that the maps in the stack share most of the structural information and, as such, there is no need to generate them separately. Thus, the last layer in the decoder outputs every texture map. Our proposed alternative architecture, $\mathcal{G}_2$ finds a shared representation of each texture map, but has a separate decoder for each of them. As such, the residual blocks are shared for all the texture stack, but each decoder can be optimized for the semantics and statistics of each particular map.

To quantitatively evaluate which architecture produces the highest quality output we compare the original texture with the generated one using standard metrics: SSIM, Si-FID, and LPIPS. The *Structural Similarity Index Measure (SSIM)* [95] is a perceptual-aware metric, working on the pixel space, that measures the similarity in structural information, and may be appropriate to evaluate synthesized textures. The Si-FID [63] is a single image extension of the *Fréchet Inception Score (FID)* [96], which measures the difference in deep latent statistics between natural and artificially generated images. Finally, we use the *Learned Perceptual Image Patch Similarity (LPIPS)* [97] as a perceptual distance metric in deep image spaces. This metric is widely used for evaluating generative models [18], [98], [99], [100], [101].

The baseline $\mathcal{G}_1$ shows more artifacts than $\mathcal{G}_2$, most likely due to the fact that the generation parts of the network are not fully separated. A quantitative evaluation is shown in Table 1, showing that $\mathcal{G}_2$ outperforms $\mathcal{G}_1$ from perceptual and statistical standpoints. Those metrics require the input and target images to have the same spatial dimensions. To obtain this, we crop the $50\%$ center area of each generated stack, which doubles the dimensions of the inputs due to the adversarial expansion approach we follow, and compare them with the input textures. In summary, $\mathcal{G}_2$ allows to generate textures that better preserve the properties of the input images without any additional computational cost and without requiring to modify the loss function, or the discriminator design.

### 7.2 Inter-map Consistency

Whilst our generators $\mathcal{G}$ can generate high-quality tileable pairs of albedo and normal maps, there is no guarantee that those maps are pixel-wise coherent, as no part in the loss function explicitly accounts for this relationship. The $\mathcal{L}_{style}$ loss is computed separately for each map in the texture stack, which may generate

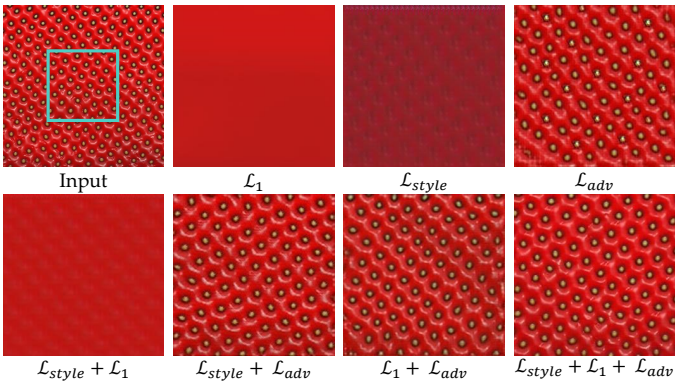



Fig. 8: Ablation study on the impact of the loss function on the quality of the synthesized textures. We evaluate each network using the same input, marked using a blue box. As shown, training $\mathcal{G}$ using the full loss function yields the best results. Each component is weighted by a $\lambda$, as specified in Section 6.

non-coherent gradients. Furthermore, our architecture of choice $\mathcal{G}_2$ separates the decoding of each map in different parts of its architecture, which may hinder the generation of spatially-coherent maps. Nevertheless, we show that this is not a problem in practice. Figure 7 shows crops of our synthesized texture maps. It can be seen that pixel-wise coherency between maps is preserved even for challenging geometric structures. We argue that the role of the discriminator is key to detect inter-layer inconsistencies by yielding lower probabilities for non-coherent maps. Even if separating the decoders may increase the risk of incoherent texture maps, this coherence is forced by the discriminator during training. To test this empirically, for the textures in Figure 7, we fed the discriminator with a crop of the original stack, a crop with the normals translated 5 px, and translated 100 px, for which we obtain average values of 0.99, 0.35, and 0.32, respectively. This suggests that, since during training, the discriminator receives the entire texture stack at once, it learns to identify spatial inconsistencies between maps.

### 7.3 Loss Function Ablation Study

As described in Section 4, the generator $\mathcal{G}$ is trained to minimize a loss function, which is comprised of adversarial, a perceptual and pixel-wise components. Each of these have different impacts on the output synthesis. We hereby study the impact of each of these components. To better isolate the impact of each metric, we perform this study using a limited generator that only outputs albedo maps. As we show on Figure 7.3, the adversarial loss $\mathcal{L}_{adv}$ provides high-level semantic consistency, the $\mathcal{L}_1$ norm acts as a regularization method which removes artifacts, while the perceptual loss $\mathcal{L}_{style}$ adds additional details to fully represent the texture. Similar findings are reported in [12]. None of these loss functions yield compelling or artifact-free results on isolation, being the adversarial loss the most important factor of the global function. Further results are included in the Supplementary Material.

## 8 Results and Comparisons

Most tileable texture synthesis methods have not shown results on synthesizing texture stacks. While adding this capability is reasonably easy for some *non parametric* methods [21], [22], others require major changes in their design. In particular, as we have shown in Section 7.1, expanding the architecture of deep learning methods for generating more than one map requires special

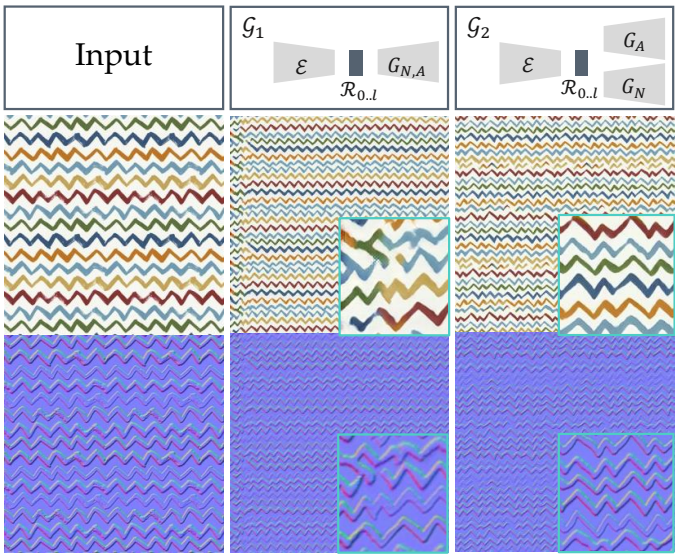



Fig. 9: A comparison of the results of our proposed architectures. Separating the decoders of the network for each map ($\mathcal{G}_2$) outperforms the joint-decoder baseline ($\mathcal{G}_1$) for both texture maps in the stacks. It generally allows for more varied albedo maps, with a more correct structure and style; as well as more accurate and sharper normal maps. The outputs in this figure are not tiled, for improved visibility of the artifacts created by $\mathcal{G}_1$. In blue, we show an up-sampled crop of the output textures. We refer the reader to the Supplementary Material for more qualitative comparisons.

attention to balance the map's interdependence with the network's capacity and expected quality. Therefore, in this section, in order to be able to compare our method with other works, we use a reduced version of our generator that only outputs a single albedo map.

**Qualitative Analysis**: First, we compare with the Texture Stationarization algorithm proposed in [13] using examples taken from their own dataset, as their code is not publicly available. A key difference between both methods is that, while their method aims to maximize the stationarity properties of the textures using external metrics, our method learns them from the texture itself, using a self-supervised approach. This results in our method better preserving the content of the original input. As shown in Figure 10, our method shows compelling results for the kind of textures shown in their paper. In the *fence* example, our methods better preserves the vertical straight lines. For the *wall* example, both methods shows compelling results, capturing a different repetition pattern. We include more results using this dataset, in addition to comparisons with the other methods on the Supplementary Material.

As Moritz *et al.*'s dataset contains mainly textures of human-made environments, we gather a different set of images of greater variety in their regularity and content. Figure 11 shows a comprehensive comparison with the other methods. The work by Li *et al.* [22], based on *Graph-Cuts*, works reasonably well if the transformation can be done locally at the seams but fails when the required changes are global, as happens in the *basket*. The *Repeated Pattern Detection* algorithm proposed by Rodriguez-Pardo *et al.* [21] is not able to handle many of these challenging cases, which do not represent grid-like textures, as in the *bananas* or *stars*. The Periodic Spatial GAN (PSGAN) proposed by Bermann *et al.* [16] generates artifacts, not maintaining the content of the texture, as in the *daisies* or the *hieroglyph* examples. A similar effect is observable in the method of [23], that use self-organizing representations through

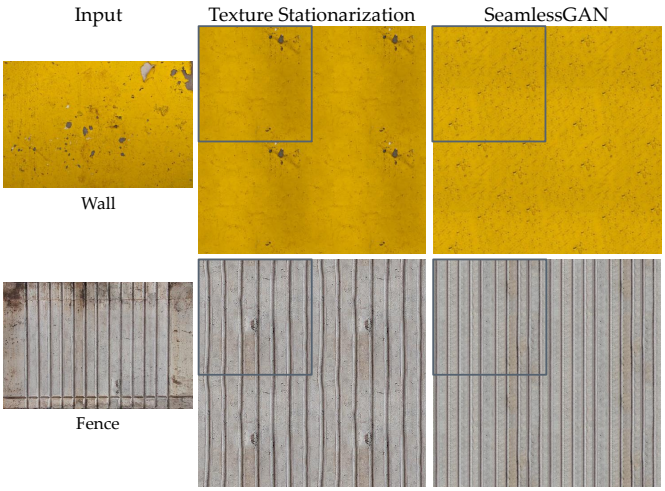



Fig. 10: A comparison with the work on Texture Stationarization by Moritz *et al.* [13], using textures from their dataset. The results are tiled $2 \times 2$ times to help visualization.

neural cellular automata. The output of the method is seamless but the integrity of the original texture is not preserved in many cases, for example, in the *hieroglyph* or the *stars*.

Our method favors keeping high-level and bigger semantic structures of the textures, resulting in larger samples with more variety. The modification applied to the texture is the minimum required to make it seamlessly tileable. It provides high-quality outputs for regular textures with uneven illumination and perspective distortion, as the *basket*, textures which are greatly irregular, such as the *bananas*, and stochastic ones, like the *strawberry*. While non-parametric synthesis methods are typically computationally cheap, obtaining results in less than one minute [20], [21], [22], at the cost of quality and generality, parametric methods are more expensive and slower. Our method is competitive in computational cost when compared to other parametric synthesis methods. Our entire training and sampling process takes less than 45 minutes. In comparison, PSGAN [16] needs 4 hours to train, the work by Zhou *et al.* [12] requires 6 hours, and the method of Niklasson *et al.* [23] takes 35 minutes to generate textures limited to $200 \times 200$ pixels. On the contrary, our method is not limited by the size of the input texture and can synthesize tiles of any size thanks to the fully-convolutional architecture, hardware being the only limiting factor. We include further and full-resolution results on the Supplementary Material.

**Quantitative Comparison**: Following a similar evaluation scheme as proposed in Section 7.1 for measuring the differences between input and synthesized images, a quantitative evaluation is shown in Table 2. We use Moritz's dataset for a fair comparison with every method. All the images we used are included in the Supplementary Material. The metrics we use for this experiment require that the input and synthesized images have the same resolution. To achieve this, and, in order to account for artifacts in the borders of the generated textures, we tile the synthesized images until they cover the same resolution as their corresponding inputs.

Interestingly, methods based on patches [13], [22], [45] obtain better SSIM and Si-FID scores than previous deep learning-based methods [16], [23]. This difference is not seen in the LPIPS metric. This suggests that patch-based methods preserve better the structure of the input textures than previous neural parametric models. Our method, by combining a variety of loss functions, allows for a better preservation of the style and semantic content of the generated textures. Furthermore, our latent space manipulation algorithm

| | SSIM ↑ | Si-FID ↓ | LPIPS ↓ |
|---|---|---|---|
| Deloit *et al.* [20] | 0.1424 | 1.3471 | 0.6207 |
| Li *et al.* [22] | 0.2086 | 0.9529 | 0.5818 |
| Moritz *et al.* [13] | 0.1968 | 0.7620 | 0.5171 |
| Rodriguez *et al.* [21] | 0.2144 | 1.2958 | 0.5137 |
| Bergmann *et al.* [16] | 0.1723 | 1.4355 | 0.5624 |
| Niklasson *et al.* [23] | 0.1753 | 1.3171 | 0.5328 |
| **SeamlessGAN** | **0.2341** | **0.6311** | **0.4792** |

TABLE 2: Quantitative comparison between different methods. We show the average results on different perceptual metrics across a variety of textures, shown in the Supplementary Material. As shown, SeamlessGAN consistently outperforms its counterparts at every studied metric. We tile the outputs until they match the spatial resolution of the input examples. Higher is better for SSIM, while lower is better for Si-FID and LPIPS.

allows for seamless borders between tiles, outperforming both previous non-parametric and parametric methods in perceptual and structural similarity. The magnitude of the Si-FID metric varies significantly between the values obtained in Table 1 and Table 2, which indicates that this metric may be overly sensitive to the global statistics of the dataset.

**Limitations and Discussion**: Our method is inherently limited by the capabilities of the adversarial expansion technique to learn the implicit structure of the given texture. That is, if the input does not show enough regularity, the adversarial expansion fails, as shown in Figure 12. It is interesting to see that the synthesis of the *dotted* texture fails to reproduce the larger dots, as they are very scarce. However, the remaining structure is very well represented.

Tiling single texture maps, even if they contain seamless borders, may generate perceptible repetitions. This is an intrinsic limitation of any single sample tileable texture synthesis. Approaches such as *Wang Tiles* [95] can tackle these limitations, but have additional disadvantages like increased memory and run-time consumption, rendering them less practical for real-time applications. Alternatively, procedural methods [3] are the most effective way to generate material samples preserving textural properties at multiple scales; however, present several limitations in the range of materials that can be generated to make them fully usable.

## 9 Conclusions and Future Work

We have proposed a deep parametric texture synthesis framework capable of synthesizing textures into tileable single-tiles, by combining recent advances on deep texture synthesis, adversarial neural networks and latent spaces manipulation. Our results show that our method can generate visually pleasing results for images with different levels of regularity and homogeneity. This work is the first method capable of exploiting properties of deep latent spaces within neural networks for generating seamless textures, and opens the opportunity for end-to-end tileable texture synthesis methods without the need for manual input. Comparisons with previous state-of-the-art methods show that our method provides results which better maintain the semantic properties of the textures, while being able to synthesize multiple maps at the same time.

Our method can be improved in several ways. First, the adversarial expansion framework, while powerful, it has some potential pitfalls that hinders its widespread applicability. The same neural architecture is used for every texture but, as discussed, different choices on the architecture make different assumptions on the nature of the textures. We have proposed a generic architecture that works well for many examples, but recent advances in Neural Architecture Search [102] may provide better priors on the optimal


Fig. 11: Comparison with previous methods. On the left column, we show the input textures. From left to right, the synthesized results of Histogram Blending, by Deloit *et al.* [20], the GraphCuts algorithm by Li *et al.* [22], Repeated Pattern Detection by Rodriguez-Pardo *et al.* [21], PSGAN by Bergmann *et al.* [16], Self-Organizing Textures by Niklasson *et al.* [23]; and ours. Outputs are tiled a similar number of times (at least twice in each dimension) for better visualization. Our method generally captures better the overall structure of the texture, while providing seamless and semantically coherent borders, for enhancing tileability. From top to bottom, we sample $c = 4, 12, 5, 17, 13, 1, 19$ and $43$ input crops before obtaining a tileable texture. More results are included in the Supplementary Material.

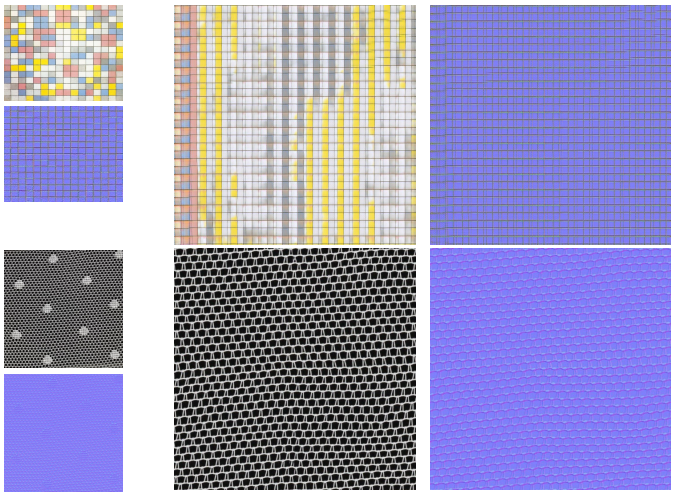


Fig. 12: A limitation of our texture synthesis algorithm: Left: Input texture stacks; right: synthesized stacks. The network fails to replicate the pattern if the occurrence is not frequent enough, as we can see in the base color of these two examples. On the other hand, the synthesized normal map is consistent as its repetitive structure is seen frequently by the network.

neural architecture to use for each sample. Furthermore, each texture synthesis network is trained from scratch. This is not only computationally costly, but learning to synthesize one texture may help in the synthesis of other textures, as shown by [15]. Fine-tuning pre-trained models for generating new textures may provide cheaper syntheses. Besides, our discriminator, while capable of detecting local artifacts, provides little control for separating such artifacts and global semantic errors. Recent work on image synthesis may provide guidance onto designing better discriminative models [88], training procedures [103], [104], image parametrizations [23], [105], [106] or perceptual loss functions [53].

Second, we proposed a synthesis solution based on manipulating latent spaces within the generative model, but explicitly training the network to generate tileable textures may provide better results than our approach. Besides, our sampling procedure could be extended for better selection of textures, by comparing histograms of the activations of the discriminator on the selected central area, instead of simply comparing minimum values.

Finally, our method has the advantage of being fully automatic, however, pre-processing the texture images so they are more easily tileable can help the synthesis process. For example, automatically rotating the textures so their repeating patterns are aligned with the axes was studied by [21]. Powerful methods for artifact [107] and distortion [108] removal could be applied as a pre-processing operation to the input textures before training the generative models, or as additional components to their loss function for improving tileability or homogeneity.

## REFERENCES


[1] P. Tu, L.-Y. Wei, K. Yatani, T. Igarashi, and M. Zwicker, "Continuous curve textures," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, pp. 1–16, 2020.

[2] Y. Hu, J. Dorsey, and H. Rushmeier, "A novel framework for inverse procedural texture modeling," *ACM Transactions on Graphics (ToG)*, vol. 38, no. 6, pp. 1–14, 2019.

[3] P. Guehl, R. Allègre, J.-M. Dischler, B. Benes, and E. Galin, "Semi-procedural textures using point process texture basis functions," in *Computer Graphics Forum*, vol. 39, no. 4, 2020, pp. 159–171.

[4] B. Galerne, A. Lagae, S. Lefebvre, and G. Drettakis, "Gabor noise by example," *ACM Transactions on Graphics (ToG)*, vol. 31, no. 4, pp. 1–9, 2012.

[5] G. Gilet, B. Sauvage, K. Vanhoey, J.-M. Dischler, and D. Ghazanfarpour, "Local random-phase noise for procedural texturing," *ACM Transactions on Graphics (ToG)*, vol. 33, no. 6, pp. 1–11, 2014.

[6] G. Guingo, B. Sauvage, J.-M. Dischler, and M.-P. Cani, "Bi-layer textures: A model for synthesis and deformation of composite textures," in *Computer Graphics Forum*, vol. 36, no. 4, 2017, pp. 111–122.

[7] E. Heitz and F. Neyret, "High-performance by-example noise using a histogram-preserving blending operator," *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, vol. 1, no. 2, pp. 1–25, 2018.

[8] Y. Guo, C. Smith, M. Hašan, K. Sunkavalli, and S. Zhao, "Materialgan: Reflectance capture using a generative svbrdf model," *ACM Transactions on Graphics (ToG)*, vol. 39, no. 6, pp. 254:1–254:13, 2020.

[9] Z. Li, K. Sunkavalli, and M. Chandraker, "Materials for masses: Svbrdf acquisition with a single mobile phone image," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 72–87.

[10] A. A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," in *Proceedings of the 28th annual conference on Computer Graphics and Interactive Techniques*, 2001, pp. 341–346.

[11] V. Kwatra, I. Essa, A. Bobick, and N. Kwatra, "Texture optimization for example-based synthesis," in *ACM Transactions on Graphics (ToG)*, vol. 24, no. 3. ACM, 2005, pp. 795–802.

[12] Y. Zhou, Z. Zhu, X. Bai, D. Lischinski, D. Cohen-Or, and H. Huang, "Non-stationary texture synthesis by adversarial expansion," *ACM Transactions on Graphics (ToG)*, vol. 37, no. 4, Jul. 2018.

[13] J. Moritz, S. James, T. S. Haines, T. Ritschel, and T. Weyrich, "Texture stationarization: Turning photos into tileable textures," in *Eurographics Symposium on Geometry Processing*, vol. 36, no. 2, 2017, pp. 177–188.

[14] A. Frühstück, I. Alhashim, and P. Wonka, "TileGAN: Synthesis of large-scale non-homogeneous textures," *ACM Transactions on Graphics (ToG)*, vol. 38, no. 4, 4 2019.

[15] G. Liu, R. Taori, T.-C. Wang, Z. Yu, S. Liu, F. A. Reda, K. Sapra, A. Tao, and B. Catanzaro, "Transposer: Universal texture synthesis using feature maps as transposed convolution filter," *arXiv preprint arXiv:2007.07243*, 2020.

[16] U. Bergmann, N. Jetchev, and R. Vollgraf, "Learning texture manifolds with the periodic spatial gan," pp. 469–477, 2017.

[17] N. Jetchev, U. Bergmann, and R. Vollgraf, "Texture synthesis with spatial generative adversarial networks," *arXiv preprint arXiv:1611.08207*, 2016.

[18] M. Mardani, G. Liu, A. Dundar, S. Liu, A. Tao, and B. Catanzaro, "Neural ffts for universal texture image synthesis," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[19] A. Hertz, R. Hanocka, R. Giryes, and D. Cohen-Or, "Deep geometric texture synthesis," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, pp. 108–1, 2020.

[20] T. Deliot and E. Heitz, "Procedural stochastic textures by tiling and blending," *GPU Zen*, vol. 2, 2019.

[21] C. Rodriguez-Pardo, S. Suja, D. Pascual, J. Lopez-Moreno, and E. Garces, "Automatic extraction and synthesis of regular repeatable patterns," *Computers and Graphics (Pergamon)*, vol. 83, pp. 33–41, 10 2019.

[22] Z. Li, M. Shafiei, R. Ramamoorthi, K. Sunkavalli, and M. Chandraker, "Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2475–2484.

[23] E. Niklasson, A. Mordvintsev, E. Randazzo, and M. Levin, "Self-organising textures," *Distill*, 2021, https://distill.pub/selforg/2021/textures.

[24] V. Deschaintre, M. Aittala, F. Durand, G. Drettakis, and A. Bousseau, "Flexible svbrdf capture with a multi-image deep network," in *Computer Graphics Forum*, vol. 38, no. 4, 2019, pp. 1–13.

[25] V. Deschaintre, G. Drettakis, and A. Bousseau, "Guided fine-tuning for large-scale material transfer," in *Computer Graphics Forum*, vol. 39, no. 4. Wiley Online Library, 2020, pp. 91–105.

[26] Y. Guo, C. Smith, M. Hašan, K. Sunkavalli, and S. Zhao, "Materialgan: reflectance capture using a generative svbrdf model," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, pp. 1–13, 2020.

[27] M. F. Cohen, J. Shade, S. Hiller, and O. Deussen, "Wang tiles for image and texture generation," *ACM Transactions on Graphics (TOG)*, vol. 22, no. 3, pp. 287–294, 2003.

[28] A. Akl, C. Yaacoub, M. Donias, J.-P. Da Costa, and C. Germain, "A survey of exemplar-based texture synthesis methods," *Computer Vision and Image Understanding*, vol. 172, pp. 12–24, 2018.

[29] L. Raad, A. Davy, A. Desolneux, and J.-M. Morel, "A survey of exemplar-based texture synthesis," *Annals of Mathematical Sciences and Applications*, vol. 3, no. 1, pp. 89–148, 2018.

[30] A. A. Efros and T. K. Leung, "Texture synthesis by non-parametric sampling," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, vol. 2. IEEE, 1999, pp. 1033–1038.

[31] V. Kwatra, A. Schödl, I. Essa, G. Turk, and A. Bobick, "Graphcut textures: image and video synthesis using graph cuts," *ACM Transactions on Graphics (ToG)*, vol. 22, no. 3, pp. 277–286, 2003.

[32] W. Dong, N. Zhou, and J.-C. Paul, "Optimized tile-based texture synthesis," in *Proceedings of Graphics Interface 2007*, 2007, pp. 249–256.

[33] J. Portilla and E. P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *International Journal of Computer Vision*, vol. 40, no. 1, pp. 49–70, 2000.

[34] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "Patchmatch: A randomized correspondence algorithm for structural image editing," *ACM Transactions on Graphics (ToG)*, vol. 28, no. 3, p. 24, 2009.

[35] C. Barnes, F.-L. Zhang, L. Lou, X. Wu, and S.-M. Hu, "Patchtable: Efficient patch queries for large datasets and applications," *ACM Transactions on Graphics (ToG)*, vol. 34, no. 4, pp. 1–10, 2015.

[36] A. Kaspar, B. Neubert, D. Lischinski, M. Pauly, and J. Kopf, "Self tuning texture optimization," in *Computer Graphics Forum*, vol. 34, no. 2, 2015, pp. 349–359.

[37] S. Darabi, E. Shechtman, C. Barnes, D. B. Goldman, and P. Sen, "Image melding: Combining inconsistent images using patch-based synthesis," *ACM Transactions on Graphics (ToG)*, vol. 31, no. 4, pp. 1–10, 2012.

[38] Y. Zhou, H. Shi, D. Lischinski, M. Gong, J. Kopf, and H. Huang, "Analysis and controlled synthesis of inhomogeneous textures," in *Computer Graphics Forum*, vol. 36, no. 2, 2017, pp. 199–212.

[39] J. S. De Bonet, "Multiresolution sampling procedure for analysis and synthesis of texture images," in *Proceedings of the 24th annual conference on Computer Graphics and Interactive Techniques*, 1997, pp. 361–368.

[40] D. J. Heeger and J. R. Bergen, "Pyramid-based texture analysis/synthesis," in *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, 1995, pp. 229–238.

[41] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2414–2423.

[42] L. A. Gatys, A. S. Ecker, M. Bethge, A. Hertzmann, and E. Shechtman, "Controlling perceptual factors in neural style transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3985–3993.

[43] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[44] J. Ren, X. Shen, Z. Lin, R. Mech, and D. J. Foran, "Personalized image aesthetics," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017, pp. 638–647.

[45] C. Rodríguez-Pardo and H. Bilen, "Personalised aesthetics with residual adapters," in *Iberian Conference on Pattern Recognition and Image Analysis*. Springer, 2019, pp. 508–520.

[46] X. Snelgrove, "High-resolution multi-scale neural texture synthesis," in *SIGGRAPH Asia 2017 Technical Briefs*, 2017, pp. 1–4.

[47] L. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 262–270.

[48] A. Dosovitskiy and T. Brox, "Generating images with perceptual similarity metrics based on deep networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 658–666.

[49] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," in *International Conference on Learning Representations*, 2018.

[50] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4401–4410.

[51] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8110–8119.

[52] A. Mordvintsev, E. Randazzo, E. Niklasson, and M. Levin, "Growing neural cellular automata," *Distill*, 2020, https://distill.pub/2020/growing-ca.

[53] E. Heitz, K. Vanhoey, T. Chambon, and L. Belcour, "A sliced wasserstein loss for neural texture synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9412–9420.

[54] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[55] A. Shocher, N. Cohen, and M. Irani, ""zero-shot" super-resolution using deep internal learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[56] Y. Asano, C. Rupprecht, and A. Vedaldi, "A critical analysis of self-supervision, or what we can learn from a single image," in *International Conference on Learning Representations*, 2019.

[57] D. Ha, A. Dai, and Q. V. Le, "Hypernetworks," *arXiv preprint arXiv:1609.09106*, 2016.

[58] V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[59] M. Tancik, P. P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. T. Barron, and R. Ng, "Fourier features let networks learn high frequency functions in low dimensional domains," *Advances in Neural Information Processing Systems*, 2020.

[60] E. Dupont, Y. W. Teh, and A. Doucet, "Generative models as distributions of functions," *arXiv preprint arXiv:2102.04776*, 2021.

[61] A. Shocher, S. Bagon, P. Isola, and M. Irani, "Ingan: Capturing and retargeting the "dna" of a natural image," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[62] S. Benaim, R. Mokady, A. Bermano, and L. Wolf, "Structural analogy from a single image pair," in *Computer Graphics Forum*, vol. 40, no. 1. Wiley Online Library, 2021, pp. 249–265.

[63] T. R. Shaham, T. Dekel, and T. Michaeli, "Singan: Learning a generative model from a single natural image," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[64] T. Hinz, M. Fisher, O. Wang, and S. Wermter, "Improved techniques for training single-image gans," in *Proceedings - 2021 IEEE Winter Conference on Applications of Computer Vision, WACV 2021*, January 2021, pp. 1300–1309.

[65] A. Shocher, S. Bagon, P. Isola, and M. Irani, "Ingan: Capturing and retargeting the" dna" of a natural image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4492–4501.

[66] V. Sushko, J. Gall, and A. Khoreva, "One-shot gan: Learning to generate samples from single images and videos," *arXiv preprint arXiv:2103.13389*, 2021.

[67] Y. Vinker, N. Zabari, and Y. Hoshen, "Training end-to-end single image generators without gans," *arXiv preprint arXiv:2004.06014*, 2020.

[68] B. Burley and W. D. A. Studios, "Physically-based shading at disney," in *ACM SIGGRAPH*, vol. 2012. vol. 2012, 2012, pp. 1–7.

[69] L. Shi, B. Li, M. Hašan, K. Sunkavalli, T. Boubekeur, R. Mech, and W. Matusik, "Match: differentiable material graphs for procedural material capture," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 6, pp. 1–15, 2020.

[70] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. S. Lempitsky, "Texture networks: Feed-forward synthesis of textures and stylized images." in *ICML*, vol. 1, no. 2, 2016, p. 4.

[71] P. Henzler, V. Deschaintre, N. J. Mitra, and T. Ritschel, "Generative modelling of brdf textures from flash images," *arXiv preprint arXiv:2102.11861*, 2021.

[72] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.

[73] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proceedings on the International Conference on Machine Learning*, vol. 30, no. 1. Citeseer, 2013, p. 3.

[74] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8789–8797.

[75] V. Deschaintre, M. Aittala, F. Durand, G. Drettakis, and A. Bousseau, "Single-image SVBRDF capture with a rendering-aware deep network," *ACM Transactions on Graphics (ToG)*, vol. 37, no. 4, 2018.

[76] H. Nam and H. E. Kim, "Batch-instance normalization for adaptively style-invariant neural networks," Tech. Rep., 2018.

[77] K. Kurach, M. Lučić, X. Zhai, M. Michalski, and S. Gelly, "A large-scale study on regularization and normalization in gans," in *International Conference on Machine Learning*. PMLR, 2019, pp. 3581–3590.

[78] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2016-Decem, 2016, pp. 770–778.

[79] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[80] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5967–5976.

[81] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, vol. 2017-Octob, pp. 2242–2251, 3 2017.

[82] M. Janner, J. Wu, T. D. Kulkarni, I. Yildirim, and J. Tenenbaum, "Self-supervised intrinsic image decomposition," in *Advances in Neural Information Processing Systems*, 2017, pp. 5936–5946.

[83] Y. Yu and W. A. Smith, "Inverserendernet: Learning single image inverse rendering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3155–3164.

[84] Z. Li and N. Snavely, "Learning intrinsic image decomposition from watching the world," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[85] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2017-Janua, pp. 105–114, 9 2017.

[86] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, vol. 3, no. January. Neural information processing systems foundation, 6 2014, pp. 2672–2680.

[87] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image Style Transfer Using Convolutional Neural Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2016-Decem, 6 2016, pp. 2414–2423.

[88] E. Schonfeld, B. Schiele, and A. Khoreva, "A u-net based discriminator for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8207–8216.

[89] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.

[90] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," Tech. Rep., 2015.

[91] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, pp. 8026–8037, 2019.

[92] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh *et al.*, "Mixed precision training," in *International Conference on Learning Representations*, 2018.

[93] S. Marcel and Y. Rodriguez, "Torchvision the machine-vision package of torch," in *Proceedings of the 18th ACM International Conference on Multimedia*, 2010, pp. 1485–1488.

[94] K. Ikeuchi, "Determining surface orientations of specular surfaces by using the photometric stereo method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 6, pp. 661–669, 1981.

[95] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[96] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *arXiv preprint arXiv:1706.08500*, 2017.

[97] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 586–595.

[98] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[99] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[100] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, "Everybody dance now," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[101] A. Almahairi, S. Rajeshwar, A. Sordoni, P. Bachman, and A. Courville, "Augmented cyclegan: Learning many-to-many mappings from unpaired data," in *International Conference on Machine Learning*. PMLR, 2018, pp. 195–204.

[102] J. Mellor, J. Turner, A. Storkey, and E. J. Crowley, "Neural architecture search without training," in *International Conference on Machine Learning*. PMLR, 2021, pp. 7588–7598.

[103] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6023–6032.

[104] A. Sinha, K. Ayush, J. Song, B. Uzkent, H. Jin, and S. Ermon, "Negative data augmentation," *arXiv preprint arXiv:2102.05113*, 2021.

[105] J. Wang, Y. Chen, R. Chakraborty, and S. X. Yu, "Orthogonal convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[106] D. A. Hudson and C. L. Zitnick, "Generative adversarial transformers," *arXiv preprint arXiv:2103.01209*, 2021.

[107] T. Dekel, T. Michaeli, M. Irani, and W. T. Freeman, "Revealing and modifying non-local variations in a single image," *ACM Transactions on Graphics (ToG)*, 2015.

[108] X. Li, B. Zhang, P. V. Sander, and J. Liao, "Blind geometric distortion correction on images through deep learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4855–4864.

**Carlos Rodriguez - Pardo** is a research engineer at SEDDI and a PhD student at the Universidad Carlos III de Madrid, Spain (UC3M). His research interests include computer vision and artificial intelligence. In 2018, he was awarded a distinction at the MSc in Artificial Intelligence at the University of Edinburgh. He completed a double BSc degree in Computer Science and Business Administration (UC3M) in 2017. He was a researcher at the Applied Artificial Intelligence Group (UC3M), working in AR applications (2013) and in data science problems (2016-2017). Carlos has served as a reviewer to conferences and journals, such as CVPR, ICCV, BMVC, ICLR, or TVCJ.

**Elena Garces** received her PhD degree in Computer Science from the University of Zaragoza in 2016. During her PhD studies, she interned twice at the Adobe (San Jose, and Seattle, USA). Her thesis dissertation focused on inverse problems of appearance capture, intrinsic decomposition from single images, video, and lightfields. She was post-doctoral researcher (2016-2018) at Technicolor R&D (Rennes, France) working on lightfields processing, and post-doctoral Juan de la Cierva Fellow (2018-2019) at the Multimodal Simulation Lab (URJC). Since 2019 she is Senior Research Scientist at SEDDI, leading the optical capture and rendering teams. She has published over 15 papers in top-tier conferences in the areas of computer graphics, vision, and machine learning, as well as authored six patents. Elena serves regularly as reviewer or PC-Member in top-tier computer vision and graphics conferences and journals such as SIGGRAPH, CVPR, ICCV, IJCV, TVCG, or EGSR.