A Patch-Based Algorithm for Diverse and High Fidelity Single Image Generation

Nicolas Cherel, Andrés Almansa, Yann Gousseau, Alasdair Newson

A PATCH-BASED ALGORITHM FOR DIVERSE AND HIGH FIDELITY SINGLE IMAGE GENERATION

Nicolas Cherel[†] *Andrés Almansa*[⋆] *Yann Gousseau*[†] *Alasdair Newson*[†]

[†]LTCI, Télécom Paris, Institut Polytechnique de Paris
[⋆]MAP5, CNRS & Université de Paris

ABSTRACT

Image generation is the task of producing new samples from one or several example images. Until recently, this has been done using large image databases, in particular using Generative Adversarial Networks (GANs). However, Shaham *et al*. [1] recently proposed the SinGAN method, which achieves this generation using a single image example. At the same time, researchers are realizing that classical patch-based methods can replace certain neural networks, with no costly training. In this paper, we present a purely patch-based method, named Patches for Single image generation (**PSin**), which requires no training and generates samples in seconds. Our algorithm is based on the minimization of a global, patch-based energy functional, which ensures the visual fidelity of the result to the original image. We also ensure diversity of the results by carefully choosing the initialization of the algorithm. We propose two initialization variants. We compare our results to both the original SinGAN and another recent patch-based image generation approach, both qualitatively and quantitatively using multiple metrics.

Index Terms— patch, single image generation, generative adversarial networks

1. INTRODUCTION

For more than twenty years, patch-based methods have represented a powerful way to generate and edit images. These methods have first been proposed in pioneering works for texture synthesis [2, 3] and then extended to other tasks such as image inpainting [4] or editing [5]. Subsequently, in the past several years, the new deep learning paradigm has taken centre stage in image processing and computer vision, and was quickly extended from classification to image generation tasks. In particular, Generative Adversarial Networks (GANs) [6] have greatly improved the capacity of these generative models to faithfully reproduce the distribution of images or of image patches. This has allowed the synthesis of new images of unprecedented quality [7]. A common characteristic of all these methods is their need for large learning databases. Recently, it has been suggested that a GAN architecture could generate new images from a single training image, through the so-called SinGAN architecture [1], confirming previous findings that the convolutional architecture is a good image prior for image restoration tasks [8].

At the same time, it has been remarked that current deep neural networks have some underlying similarities with patch-based methods (eg. in the case of attention mechanisms). Patch-based methods, however, require none of the costly training entailed by neural networks. Consequently, Granot *et al*. [9] proposed a relatively simple patch-based method, free of deep convolutional architecture, to challenge SinGAN's approach to the generation of new image content and yield better visual quality at a reduced computing time.

The main two goals of image synthesis are to produce results with both high visual fidelity with respect to the original image, but that also have enough diversity. Indeed, it is trivial to achieve fidelity by always producing the same image, and conversely it is trivial to produce high diversity by producing noise. Thus, attaining both goals is a great challenge.

In this work, we propose an efficient, fully patch-based method for single image synthesis, requiring no training, which produces examples of both high fidelity and diversity. We encourage fidelity by minimizing a patch-based energy in a multi-scale approach, and we ensure diversity by carefully choosing the initialization of the example, which turns out to be crucial for diversity. In particular, we propose an initialization based on optimal transport, designed to respect the patch distribution of the original image. We show that our approach produces better results than the original SinGAN and the other patch-based method of Granot *et al*., which has difficulty in ensuring diversity. Furthermore, our method is computationally fast, generating samples in seconds.

In Section 2, we review some of the previous work related to this subject, in Section 3 we present our image generation method, and finally we present our results in Section 4.

2. RELATED WORK

Texture synthesis. Efros and Leung [2] were among the first to propose the use of image patches for synthesis purposes, with a greedy non-parametric algorithm for texture synthesis. Kwatra *et al*. [3] optimized a global energy, minimized

To appear in *Proc. ICIP2022, October 16-19, 2022, Bordeaux, France*

Algorithm 1 PSin / PSinOT

$u \leftarrow$ init() ▷ Random noise or optimal initialization
for $s \in [2, 1, 0]$ **do**
  $u \leftarrow$ rescale$(u, \text{scale} = s)$
  for $i = 1..10$ **do**
    $\phi \leftarrow$ NN-Mapping$(u, \tilde{u})$ ▷ Expectation
    $u \leftarrow$ Reconstruction$(\phi, \tilde{u})$ ▷ Maximization
  end for
end for

at different scales. Wexler *et al*. [4] proposed a similar energy for video inpainting. The idea was then improved upon by Newson *et al*. [10] who sped up the technique by adapting PatchMatch [11] to video. Recently, Gatys *et al*. [12] have generated textures by an optimization approach. They used a pretrained classification network and matched deep feature statistics for texture generation. Ulyanov *et al*. [13] trained a network to generate images from noise minimizing the same criterion. Based on optimal transport, Houdard *et al*. [14] proposed a multiscale patch-based texture synthesis which respects the patch distribution of the reference image.

Single Image Learning. Deep Image Prior by Ulyanov *et al*. [8] has shown that deep convolutional networks are very good priors for image processing tasks. An untrained network was used as a regularization term in the objective function. Shaham *et al*. [1] have proposed SinGAN for single image generation. SinGAN uses a pyramid and a simple patch generator / discriminator architecture at each level for generating variations of a single image. InGAN [15] was proposed by Shocher *et al*. for learning the internal patch distribution of an image. Once trained, the network can resize an image while preserving the patch distribution. Granot *et al*. [9] very recently proposed an approach which is similar to ours in spirit, showing that GANs are indeed not necessary for single image generation. They addressed the problem of distribution with a hand-designed regularization term.

3. PATCHES FOR SINGLE IMAGE GENERATION

We propose **PSin**, a **P**atch-based algorithm for **Sin**gle image generation. Our algorithm exploits the gaussian pyramid, minimizing an energy from the coarsest scale to the finest scale. We avoid costly learning stages by copying patches from the reference image.

We now introduce the patch-based optimization problem that we solve to produce the output image. Let $\tilde{u}$ be the reference image and u be the new, synthesised image, defined over the image domain Ω. A patch centred on a pixel p in image u is denoted as Ψ_p^u. At each scale, we minimize a global energy similar to the one of Wexler *et al*. [4] or Kwatra *et al*. [3]:

$$E(u) = \sum_{p \in \Omega} \min_{\tilde{p} \in \Omega} \|\Psi_p^u - \Psi_{\tilde{p}}^{\tilde{u}}\|_2^2, \quad (1)$$

Algorithm	Learning-Free	Distribution	Scalable
SinGAN [1]	✗	✓	✓
GPNN [9]	✓	✓	✗
PSin	✓	✗	✓
PSinOT	✓	✓	✓

Table 1. Combining PSin with a good initialization gives an algorithm that does not require learning, respects the original distribution, scales to higher images and has limited runtime

where $\|\Psi_p^u - \Psi_{\tilde{p}}^{\tilde{u}}\|_2^2$ is the ℓ_2 distance between the pixels of the patches Ψ_p^u and $\Psi_{\tilde{p}}^{\tilde{u}}$. This energy specifies that a good solution is one where each patch is similar to its nearest neighbor (NN), with respect to the ℓ^2 patch distance, in the reference image.

This energy is efficiently minimized by alternating a nearest neighbor search step and a reconstruction step, which have been identified as two steps of a hard Expectation Maximization (EM) by Kwatra *et al*. [3]. Let $\phi(p)$ represent the nearest neighbours mapping, in other words $\phi(p) = \arg\min_{\tilde{q} \in \Omega} \|\Psi_p^u - \Psi_{\tilde{q}}^{\tilde{u}}\|_2^2$. The reconstruction step is given, for each pixel p by:

$$u(p) = \sum_{q \in \Psi_p} e^{-\|\Psi_q^u - \Psi_{\phi(q)}^{\tilde{u}}\|_2^2} \, \tilde{u}(\phi(q) + (q - p)) \quad (2)$$

Using the efficient approximate nearest neighbor algorithm PatchMatch [11] makes generation possible in seconds. No training is required at any time. Our full algorithm is described in Alg 1.

This energy is minimized at each scale starting from the coarsest to the finest scale, adding more and more details. The coarse structure *e.g*. position of the main objects and structures, is defined at the very beginning similarly to SinGAN. The initialization and first scales are thus crucial steps in our algorithm, and must be carefully considered. We have two strategies for initialization.

3.1. Random Initialization - PSin

In the simplest approach that we first consider, Gaussian noise can be used as an initialization. We refer to this method as **PSin**. While simple, this can lead to interesting structures provided that the starting resolution is low enough with respect to the patch size. Figure 1 illustrates this phenomenon: with 3 scales, the generated image has poor global coherency. In general, this approach ensures some diversity but has limited fidelity, in the sense that *e.g*. it does not respect the distribution of patches in the reference image. To address this problem, we turn to another initialization.

3.2. Optimal distribution- PSinOT

In this approach, we turn to tools from optimal transport to build a loss that accounts for the distance between the proba-

Fig. 1. PSin results with 3 scales (left) vs 4 scales (right). With 3 scales, the general structure is not coherent; this is addressed by initializing at a lower scale.

bility distribution of patches from the input and the one of the synthesized image. This enables us to produce an initialization which has a similar patch distribution to the reference image. This loss, the Wasserstein-2 distance, is minimized at a coarse scale to produce the desired initialization. Our method is inspired by the work of Houdard *et al*. [14], who proposed a patch-based optimal transport algorithm for texture synthesis. Minimizing the Wasserstein-2 distance ensures that the generated samples have the correct patch distribution, ie the distribution of patches in the reference image.

Using our notations, the semi-dual problem at a single scale is the following:

$$OT(u) = \max_{\beta} \sum_{p\in\Omega} \min_{\tilde{p}\in\Omega} \left(\|\Psi_p^u - \Psi_{\tilde{p}}^{\tilde{u}}\|_2^2 - \beta_{\tilde{p}} \right) + \sum_{\tilde{p}\in\Omega} \beta_{\tilde{p}} \quad (3)$$

using $\beta \in \mathbb{R}^{|\Omega|}$ as the dual variable. The cost is minimized by alternative optimizations on u and β.

This process is long and computationally expensive. It scales quadratically with the number of patches which makes it unpractical for single image generation. To combine the strengths of both approaches, we propose to first create a coarse initialization by optimization and then switch to our fast generator for performance. We call this method **PSinOT**.

3.3. Fast nearest neighbor search

Our algorithm spends most of its computational effort finding the nearest neighbors. Unfortunately, a naive approach to this search does not scale well, with a complexity of $\mathcal{O}(n^2)$ for n patches. Therefore, we turn to PatchMatch [11] for a fast computation of nearest neighbors. This makes PSin and PSinOT scalable algorithms. Table 1 summarizes the advantages of each method. In practice, PSin can generate a new sample in 15 seconds on the CPU, while SinGAN first requires 1 hour of training on GPU. GPNN takes 6 seconds to generate a sample on GPU. The optimal initialization in PSinOT adds 15s (see Table 2 for running times).

Algorithm	Runtime CPU (s)	Runtime GPU (s)
SinGAN	-	3700
GPNN	32	6
PSin	15	-
PSinOT	72	30*

Table 2. Runtimes of each algorithm on CPU / GPU. PSin is the fastest on CPU. The optimal initialization slows down the total run time. *Initialization on GPU and refinement on CPU

4. RESULTS

4.1. Implementation details

We implement our algorithm in Python and speed it up with Numba on CPU. We typically use 4 scales with a factor 2 and set the patch size to 11, which is comparable to the receptive field of SinGAN [1]. We do 10 iterations of EM at each scale before switching. Upscaling is done by interpolating the shift map ϕ rather than interpolating the image u. For PSin, we use a gaussian noise $\mathcal{N}(0.5, 1)$. For PSinOT, optimal transport is employed for the first two scales and then PSin is used. Our patch distance includes the RGB difference and the norm of the horizontal and vertical gradients, which have been found to improve the synthesis of textures [10]. For comparisons, we have used the official implementation of each work with their default parameters[1]. ($\sigma = 0.75$ for GPNN).

4.2. Quantitative results

Evaluating image generation is challenging in itself. Therefore we rely on several metrics to compare the methods. We use the Fréchet Inception Distance [16] which measures the distance between gaussian distributions of features and its adaptation to single image generation (SIFID) [1]. A low SIFID means that images have the same feature distribution and contain the same visual objects. For fidelity, we include the optimal transport cost on patches (derived from the work of Houdard *et al*. [14]) which measures the true distance between patch distributions at the finest scale. In practice this is done by optimizing Equation 3 in the dual variable β only, with 1000 iterations of gradient ascent. We also measure the diversity of generated images. The entropy of the distribution is intractable and we use the measure of diversity given in Shaham *et al*. [1]. For each image, the pixel diversity is the standard deviation of pixel-wise intensities when stacking all generated images. We compare with the results of SinGAN, GPNN [9], PSin, and PSinOT.

We use a dataset of 50 images from Places [17], the same as in Shaham *et al*., and compute our metrics on 50 samples for each image. Table 3 confirms that PSin produces very variable results with limited fidelity. SinGAN produces

[1]GPNN: https://github.com/iyttor/GPNN, SinGAN: https://github.com/tamarott/SinGAN

Fig. 2. Reference image and 6 uncurated samples for each algorithm to showcase diversity. SinGAN and PSin produce diverse shapes, however the visual quality of SinGAN is clearly lacking. GPNN introduces very little diversity, keeping the main structure of the reference image (the single rock arch). PSinOT produces original geometries while maintaining better visual fidelity than SinGAN.

Algorithm	SIFID ↓	Optimal Transport ↓	Diversity ↑
SinGAN	0.12	1.34	0.34
GPNN	**0.02**	*0.52*	0.40
PSin	0.45	0.94	**0.62**
PSinOT	*0.06*	**0.36**	*0.53*

Table 3. PSin is very diverse but has limited fidelity (SIFID, optimal transport). PSinOT combines high diversity and similar distribution. **best**, *second best*.

Fig. 3. SinGAN (left) and PSinOT (right). Our method does not have network artifacts.

less diverse output with a lower SIFID. Finally GPNN and PSinOT both have good diversity and fidelity scores but our approaches yield significant improvements both in the fidelity of patches distribution and in diversity.

4.3. Qualitative results

We also present visual results in Figure 2 which are representative outputs. SinGAN's results are visually pleasing from a distance but suffer from network artifacts when looked at closely (Figure 3). GPNN produces visually coherent results but may reproduce the same image multiple times. Finally PSinOT has a coherent structure and satisfying details. Our website contains more examples: link to website.

5. CONCLUSION

We have presented a patch-based approach to single image generation. Contrary to SinGAN, it does not require training but still generates diverse and visually pleasing images. Our algorithm is based on the minimization of a patch energy, which encourages fidelity to the reference image. In order to ensure that the patch distribution of the reference image is respected, we propose an initialization based on optimal transport. We have compared our results quantitatively and qualitatively with the original SinGAN and another patch-based method, showing that our approach achieves better fidelity and diversity than the previous two. Although patch-based methods work well here, in future work, we may want to pre-

serve some of the advantages of convolutional architectures, replacing only the discriminator with a patch-based approach. More generally, we will investigate how patches can be included in general-purpose GANs, not just for single images.

6. REFERENCES

[1] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli, "Singan: Learning a generative model from a single natural image," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4570–4580.

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

[3] Vivek Kwatra, Irfan Essa, Aaron Bobick, and Nipun Kwatra, "Texture optimization for example-based synthesis," in *ACM SIGGRAPH 2005 Papers*, pp. 795–802. 2005.

[4] Yonatan Wexler, Eli Shechtman, and Michal Irani, "Space-Time Completion of Video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 463–476, Mar. 2007.

[5] Yael Pritch, Eitam Kav-Venaki, and Shmuel Peleg, "Shift-map image editing," in *2009 IEEE 12th international conference on computer vision*. IEEE, 2009, pp. 151–158.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Eds. 2014, vol. 27, Curran Associates, Inc.

[7] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8110–8119.

[8] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky, "Deep image prior," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[9] Niv Granot, Ben Feinstein, Assaf Shocher, Shai Bagon, and Michal Irani, "Drop the gan: In defense of patches nearest neighbors as single image generative models," *arXiv preprint arXiv:2103.15545*, 2021.

[10] Alasdair Newson, Andrés Almansa, Matthieu Fradet, Yann Gousseau, and Patrick Pérez, "Video Inpainting of Complex Scenes," *SIAM Journal on Imaging Sciences*, vol. 7, no. 4, pp. 1993–2019, 2014.

[11] Connelly Barnes, E. Shechtman, A. Finkelstein, and Dan B. Goldman, "PatchMatch: a randomized correspondence algorithm for structural image editing," in *SIGGRAPH 2009*, 2009.

[12] Leon Gatys, Alexander S Ecker, and Matthias Bethge, "Texture synthesis using convolutional neural networks," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds. 2015, vol. 28, Curran Associates, Inc.

[13] Dmitry Ulyanov, Vadim Lebedev, Andrea Vedaldi, and Victor S Lempitsky, "Texture networks: Feed-forward synthesis of textures and stylized images.," in *ICML*, 2016, vol. 1, p. 4.

[14] Antoine Houdard, Arthur Leclaire, Nicolas Papadakis, and Julien Rabin, "Wasserstein Generative Models for Patch-based Texture Synthesis," in *Scale Space and Variational Methods in Computer Vision*, Cabourg, France, May 2021, vol. LNCS 12679, pp. 269–280.

[15] Assaf Shocher, Shai Bagon, Phillip Isola, and Michal Irani, "InGAN: Capturing and Remapping the "DNA" of a Natural Image," *arXiv:1812.00231 [cs]*, Apr. 2019, arXiv: 1812.00231.

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[17] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba, "Places: A 10 million image database for scene recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.