SinGAN : Learning a Generative Model from a Single Natural Image

Tristan Lecourtois [1] Mathéo Kina [1]

[1]ENS Paris-Saclay

Abstract

This project presents the implementation and enhancement of **SinGAN**, a GAN trained on a single image. We replicated the original model, evaluated hyperparameter and padding choices using SIFID, and explored three alternative discriminator loss functions. Results show that fine-tuning hyperparameters improves image quality and that alternative losses can enhance efficiency while maintaining visual performance.

Original SinGAN implementation

GANs are powerful for image synthesis but usually need large datasets. Unlike GANs, the SinGAN are able to learn the structure of an image from a unique sample. Indeed, extracting patches of different scales allows SinGANs to discover the internal statistics of patches in an image.

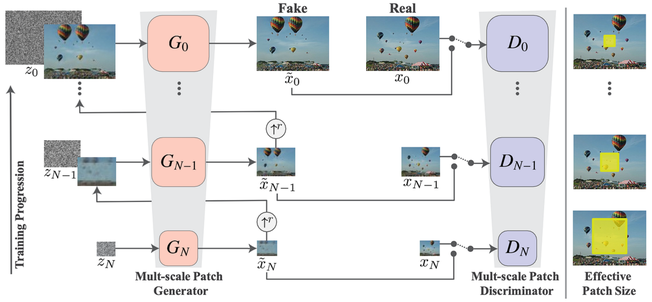

Figure 1. SinGAN multi-scale architecture. Each generator G_n refines the image progressively.

Generators: Produce images at each scale n by refining the previous upsampled image with noise z_n:

$$\tilde{x}_n = G_n(z_n, (\tilde{x}_{n+1}) \uparrow r)$$

Discriminators: Distinguish real vs generated patches at each scale:

$$L_{\text{adv}} = -\mathbb{E}[D_n(x_n)] + \mathbb{E}[D_n(\tilde{x}_n)] + \lambda \mathbb{E}[(\|\nabla D_n(x)\|_2 - 1)^2]$$

Reconstruction loss: Ensures existence of a noise map reproducing the original image:

$$L_{\text{rec}} = \|\tilde{x}_n - x_n\|_2^2$$

Training objective:

$$\min_{G_n} \max_{D_n} L_{\text{adv}}(G_n, D_n) + \alpha L_{\text{rec}}(G_n)$$

Pre-processing

Input images are standardized to the range $[-1, 1]$ to stabilize GAN optimization and ensure consistent input scales across generators. Images are resized using scale-aware interpolation with fixed output sizes, guaranteeing shape consistency across downsampling and upsampling steps. Finally, the number of scales and scaling factor are automatically computed so that the coarsest image reaches a minimum resolution, enabling effective multi-resolution learning.

Training Results

Evaluation Metric. The Single Image Fréchet Inception Distance (SIFID) is used to quantitatively measure the similarity between real and generated images at each scale:

$$\text{SIFID}(x, \tilde{x}) = \|\mu_x - \mu_{\tilde{x}}\|_2^2 + \text{Tr}\left(\Sigma_x + \Sigma_{\tilde{x}} - 2(\Sigma_x \Sigma_{\tilde{x}})^{1/2}\right)$$

where (μ_x, Σ_x) and $(\mu_{\tilde{x}}, \Sigma_{\tilde{x}})$ are the mean and covariance of Inception features. At each scale, the generated image is directly compared to the corresponding downsampled real image.

Figure 2. Comparison of Real and Fake Images at Each Scale

Generating Images

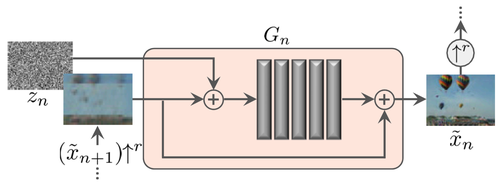

Figure 3. Single scale generation

After training on a single image, SinGAN generates diverse and high-quality samples through a multi-scale process that mirrors training: at each scale n, the image from the previous scale $\tilde{x}_{n+1}$ is upsampled, combined with a random noise map z_n, and passed through the generator to produce a residual added to the upsampled image. Formally, the generation starts from an empty image at the coarsest scale with $\tilde{x}_N = G_N(z_N)$, and proceeds as $\tilde{x}_n = G_n\left(z_n, (\tilde{x}_{n+1})^{\uparrow r}\right)$ for all $n < N$, allowing SinGAN to preserve the global structure and fine textures of the original image while ensuring sample diversity.

Figure 4. Generated Samples

Change of boundary conditions

We investigated the impact of replacing zero-padding with circular-padding in the generator, as the original paper provides no explicit justification for the padding choice. While circular-padding preserved local visual elements, it disrupted global spatial structure, leading to misplaced objects and incoherent scene layouts

Discriminator Loss Experiments

Explored three alternatives:

1. **Fréchet Distance Loss**: improves efficiency, captures patch statistics

   $$\mathcal{L}_{\text{FD}} = \|\mu_r - \mu_g\|_2^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}\right)$$

   where (μ_r, Σ_r) and (μ_g, Σ_g) are the mean and covariance of real and generated patch features.
2. **Nearest Neighbor Patch Loss**: maintains visual quality

   $$\mathcal{L}_{\text{NN}} = \mathbb{E}_{p_g \in P_G}\left[\min_{p_r \in P_R} \|p_g - p_r\|_2^2\right]$$

   enforcing similarity between each generated patch p_g and its closest real patch p_r.
3. **Aligned NN Patch Loss**: further enhances fine-scale consistency

   $$\mathcal{L}_{\text{Aligned-NN}} = \mathbb{E}_i\left[\|P_G(i) - P_R(i)\|_2^2\right]$$

   where patches are matched at identical spatial locations i across scales.

Substituting the network architecture with explicit losses reduces training cost per scale and keeps output quality high.

Figure 5. Comparison of Generated Images Across Scales: Real vs Frechet vs NNPL

References

Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model from a single natural image, 2019. URL `https://arxiv.org/abs/1905.01164`.