SinGAN: Learning a Generative Model from a Single Natural Image

Tamar Rott Shaham, Tali Dekel, Tomer Michaeli

Presented by: Yuzeng You
Outline

• Introduction
• Method
• Results
• Applications
• Conclusion
Introduction

- SinGAN
  - unconditional generative model
  - learn from a single natural image
Introduction

• Related work
  • Single image deep models
  • Generative models for image manipulation
Figure 3: **SinGAN vs. Single Image Texture Generation.** Single image models for texture generation [3, 16] are not designed to deal with natural images. Our model can produce realistic image samples that consist of complex textures and non-repetitive global structures.
Method

- training samples: patches of a single image
- a hierarchy of patch-GANs
Figure 4: **SinGAN’s multi-scale pipeline.** Our model consists of a pyramid of GANs, where both training and inference are done in a coarse-to-fine fashion. At each scale, $G_n$ learns to generate image samples in which all the overlapping patches cannot be distinguished from the patches in the down-sampled training image, $x_n$, by the discriminator $D_n$; the effective patch size decreases as we go up the pyramid (marked in yellow on the original image for illustration). The input to $G_n$ is a random noise image $z_n$, and the generated image from the previous scale $\tilde{x}_n$, upscaled to the current resolution (except for the coarsest level which is purely generative). The generation process at level $n$ involves all generators $\{G_N \ldots G_n\}$ and all noise maps $\{z_N, \ldots, z_n\}$ up to this level. See more details at Sec. 2.
Method

- fully convolutional net
- 5 cone-blocks
- Conv(3x3)-BatchNorm-LeakyReLU

Figure 5: Single scale generation. At each scale $n$, the image from the previous scale, $\tilde{x}_{n+1}$, is upscaled and added to the input noise map, $z_n$. The result is fed into 5 conv layers, whose output is a residual image that is added back to $(\tilde{x}_{n+1})^↑r$. This is the output $\tilde{x}_n$ of $G_n$. 
Method

• Training
  \[ \min_{G_n} \max_{D_n} \mathcal{L}_{adv}(G_n, D_n) + \alpha \mathcal{L}_{rec}(G_n). \]

• Adversarial loss: WGAN-GP

• Reconstruction loss: \[ \mathcal{L}_{rec} = \|G_n(0, (\tilde{x}_{n+1}^{\text{rec}}) \uparrow^r) - x_n \|^2, \]
Results

• Resources: Berkeley Segmentation Database (BSD), Places and the Web

• coarsest scale: 25px

• scaling factor $r$: as close as possible to $4/3$
Figure 1: **Image generation learned from a single training image.** We propose *SingGAN*—a new unconditional generative model trained on a *single natural image*. Our model learns the image’s patch statistics across multiple scales, using a dedicated multi-scale adversarial training scheme; it can then be used to generate new realistic image samples that preserve the original patch distribution while creating new object configurations and structures.
Figure 6: **Random image samples.** After training SinGAN on a single image, our model can generate realistic random image samples that depict new structures and object configurations, yet preserve the patch distribution of the training image. Because our model is fully convolutional, the generated images may have arbitrary sizes and aspect ratios. Note that our goal is not image retargeting – our image samples are random and optimized to maintain the patch statistics, rather than preserving salient objects. See SM for more results and qualitative comparison to image retargeting methods.
Figure 7: **High resolution image generation.** A random sample produced by our model, trained on the $243 \times 1024$ image (upper right corner); new global structures as well as fine details are realistically generated. See 4Mpix examples in SM.
Results

- effect of scales at test time
Results

- effect of scales during training

Figure 9: The effect of training with a different number of scales. The number of scales in SinGAN’s architecture strongly influences the results. A model with a small number of scales only captures textures. As the number of scales increases, SinGAN manages to capture larger structures as well as the global arrangement of objects in the scene.
Results

- Quantitative Evaluation
- AMT perceptual study

<table>
<thead>
<tr>
<th>1st Scale</th>
<th>Diversity</th>
<th>Survey</th>
<th>Confusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td>0.5</td>
<td>paired</td>
<td>21.45% ± 1.5%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>unpaired</td>
<td>42.9% ± 0.9%</td>
</tr>
<tr>
<td>$N - 1$</td>
<td>0.35</td>
<td>paired</td>
<td>30.45% ± 1.5%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>unpaired</td>
<td>47.04% ± 0.8%</td>
</tr>
</tbody>
</table>

Table 1: “Real/Fake” AMT test. We report confusion rates for two generation processes: Starting from the coarsest scale $N$ (producing samples with large diversity), and starting from the second coarsest scale $N - 1$ (preserving the global structure of the original image). In each case, we performed both a paired study (real-vs.-fake image pairs are shown), and an unpaired one (either fake or real image is shown). The variance was estimated by bootstrap [14].
Results

• Quantitative Evaluation

• Single Image Fréchet Inception Distance

<table>
<thead>
<tr>
<th>1st Scale</th>
<th>SIFID</th>
<th>Survey</th>
<th>SIFID/AMT Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td>0.09</td>
<td>paired</td>
<td>$-0.55$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>unpaired</td>
<td>$-0.22$</td>
</tr>
<tr>
<td>$N - 1$</td>
<td>0.05</td>
<td>paired</td>
<td>$-0.56$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>unpaired</td>
<td>$-0.34$</td>
</tr>
</tbody>
</table>

Table 2: Single Image FID (SIFID). We adapt the FID metric to a single image and report the average score for 50 images, for full generation (first row), and starting from the second coarsest scale (second row). Correlation with AMT results shows SIFID highly agrees with human ranking.
Applications

• Super-Resolution
• Paint-to-Image
• Harmonization
• Editing
• Single Image Animation
Figure 10: **Super-Resolution.** When SinGAN is trained on a low resolution image, we are able to super resolve. This is done by iteratively upsampling the image and feeding it to SinGAN’s finest scale generator. As can be seen, SinGAN’s visual quality is better than the SOTA internal methods ZSSR [46] and DIP [51]. It is also better than EDSR [32] and comparable to SRGAN [30], external methods trained on large collections. Corresponding PSNR and NIQE [40] are shown in parentheses.
Figure 11: **Paint-to-Image.** We train SinGAN on a target image and inject a downsampled version of the paint into one of the coarse levels at test time. Our generated images preserve the layout and general structure of the clipart while generating realistic texture and fine details that match the training image. Well-known style transfer methods [17, 38] fail in this task.
Figure 12: **Editing.** We copy and paste a few patches from the original image (a), and input a downsampled version of the edited image (b) to an intermediate level of our model (pretrained on (a)). In the generated image (d), these local edits are translated into coherent and photo-realistic structures. (c) comparison to Photoshop content aware move.

Figure 13: **Harmonization.** Our model is able to preserve the structure of the pasted object, while adjusting its appearance and texture. The dedicated harmonization method [34] overly blends the object with the background.
Conclusion

• SinGAN: a new unconditional generative scheme that is learned from a single natural image

• it can provide a very powerful tool for a wide range of image manipulation tasks
https://www.youtube.com/watch?v=xk8bWLZk4DU&feature=youtu.be
Thank you.
Q&A