SinGAN: Learning a Generative Model from a Single Natural Image

Tamar Rott Shaham, Tali Dekel, Tomer Michaeli

Best Paper Award (Marr Prize) in ICCV’19
Overview

• Problem Statement
• Related Work
• SinGAN
• Architecture
• Losses
• Effects of scale N
• Results
Problem Statement

• Unconditional generative model for image synthesis.
• Learning internal distribution and statistics of a natural image.
• A single model for all image manipulation tasks.
Image Manipulations

In image synthesis, image manipulation tasks are (not limited) to:

- Super Resolution
- Paint to image
- Editing
- Harmonization
- Animation
Super Resolution

Enhances the resolution of image, recovering the finer texture details at large upscaling factors.

Fig 1: https://github.com/tensorlayer/SRGAN
Paint-to-Image

Style transfer technique using textural information learned from a natural image, converting a clipart to a realistic image.
Harmonization

Copying an element into an image, with local harmonization of color, texture, and structure properties of the image.

Editing

Produce a seamless image composite in which image regions have been copied and pasted in other locations.
Animation

Creating a short video clip with realistic object motion, from a single input image.
Related Work: Deep Image Prior

• Single image-based technique.
• Relies on randomly initialized generator network that can be used as a hand-crafted prior.

$$\theta^* = \arg\min_{\theta} E(f_{\theta}(z); x_0), \quad x^* = f_{\theta^*}(z).$$

Slide No: 10
Related Work: Deep Image Prior

- Denoising Image

- Inpainting (Text, Masked region)
SinGAN: Architecture

- Sequentially trained from coarsest to the finest.
- $r>1$, $x_n$ is down sampled version of $x$ by $r_n$. 
SinGAN (Single Image GAN)

- Captures global properties including arrangement of shape, finer details and texture details.
- At each layer, patch-GANs captures the patch distribution at different scales, from coarsest to finest.
- Each Generator is coupled with a Markovian patch discriminator.
SinGAN

- Generator and Discriminator network has 5 Conv block.
- Each block has 32 kernel, 3x3 Conv | Batchnorm | LeakyReLU.
- Patch size is fixed in all the discriminator (11x11).
SinGAN: Effects of scale N

- No. of scales ~ global features and arrangements captured.
SinGAN : Effects of Injection

- We can control the generation by starting from different levels which will preserve shape and poses.
**SinGAN : Losses**

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

- Adversarial loss is for each patch calculated using WGAN-GP.
- Loss is calculated for whole image to learn boundary conditions.
- Reconstruction Loss maps random noise that generates the desired image. Determines the S.D. of the noise \( z_n \) at each scale.

\[
\text{for } n < N, \mathcal{L}_{\text{rec}} = \| G_n(0, (\tilde{z}_{n+1}^{\text{rec}}) \uparrow^r) - x_n \|^2,
\]

\[
\text{for } n = N, \mathcal{L}_{\text{rec}} = \| G_N(z^*) - x_N \|^2.
\]
**FID (Frechet Inception Distance)**

- Deviation b/w features in the real\((x)\) and the generated\((g)\) image.
- Lower FID values mean better image quality and diversity.
- Features from last pooling layer are collected from inception net.

\[
FID(x, g) = \|\mu_x - \mu_g\|^2 + \text{Tr}(\Sigma_x + \Sigma_g - 2(\Sigma_x \Sigma_g)^{\frac{1}{2}})
\]
SIFID (Single Image FID)

- Measures the internal patch statistics of an image.

- Output of the convolutional layer just before the second pooling layer between the features in the real and generated sample.
SinGAN: Image Synthesis

- Maintains the global structure/feature.
- Aspect ratio/pixel size of noise generates variety of samples.

Training image | Random samples from a single image
Dataset: Places

- 205 categories
- 7 billion images for Scene Recognition
- 50 images from categories Mountains, Hills, Desert, Sky.

*Fig*: Learning Deep Features for Scene Recognition using Places Database
Results: Quantitative

Amazon Mechanical Turk perceptual study | presented for 1 sec
• Paired (real vs. fake)
• Unpaired (either real or fake)

<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>
<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>

SIFID/AMT (anti)correlation indicates small SIFID ~ large confusion

Correlation is stronger for the paired, since SIFID is a paired measure
SinGAN : Super Resolution

- To increase the resolution by a factor s, the $L_{\text{rec}}$ weight of $\alpha = 100$ and pyramid scale factor of $r = k\sqrt{s}$ for some $k \in \mathbb{N}$.

$$\min_{G_n} \max_{D_n} \mathcal{L}_{\text{adv}}(G_n, D_n) + \alpha \mathcal{L}_{\text{rec}}(G_n).$$

- Repeat $G_0$ $k$ times to create a high definition output.

*From 243 × 1024 image to 4 Mpix image*
Dataset: BSD100

For Super Resolution task

- 200 training | 100 test images for Segmentation
- Segmentation from 3 different people
- Has probability distributions with Gestalt grouping factors
Metric : NIQE

- Blind Image Quality Assessment metric.
- Based on measurable deviations from statistical features from a corpus of natural, undistorted images.

Results : Quantitative

Natural Image Quality Evaluator (NIQE) is a perceptual quality indicator. Lower is better.

<table>
<thead>
<tr>
<th>External methods</th>
<th>Internal methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRGAN</td>
<td>EDSR</td>
</tr>
<tr>
<td>RMSE</td>
<td>16.34</td>
</tr>
<tr>
<td>NIQE</td>
<td>3.41</td>
</tr>
<tr>
<td>EDSR</td>
<td>12.29</td>
</tr>
<tr>
<td>RMSE</td>
<td>13.82</td>
</tr>
<tr>
<td>NIQE</td>
<td>6.50</td>
</tr>
<tr>
<td>DIP</td>
<td>13.08</td>
</tr>
<tr>
<td>RMSE</td>
<td>6.35</td>
</tr>
<tr>
<td>NIQE</td>
<td>7.13</td>
</tr>
<tr>
<td>ZSSR</td>
<td>16.22</td>
</tr>
<tr>
<td>SinGAN</td>
<td>3.71</td>
</tr>
</tbody>
</table>

RMSE/PSNR - measure the quality of reconstruction
SinGAN : Editing

• After training in the background image, inject the edited image to coarse level N-1 level after down sampling the edited image.
SinGAN : Paint-to-Image

• After training in the background image, inject the edited/natively pasted composite to coarse levels N-1 or N-2.
SinGAN : Harmonization

• The model can preserve the structure of the pasted object, while adjusting its appearance and texture.
• Scales 2,3,4 preserves good harmonization.
SinGAN : Video Generation

• Generating images from multiple random noise samples
• Combine synthesised frames to create realistic object motion.
Drawbacks

• Learning is inherently limited in terms of semantic diversity of the input image.
• Model will not hallucinate out of the box features/textures.
• Complexity of the generator(no of scales) is proportional to the complexity of the image being processed.
References

Questions?

Thanks :)