Artificial Fingerprinting for Generative Models: Rooting Deepfake Attribution in Training Data
Ning Yu*1,2 Vladislav Skripniuk*3 Sahar Abdelnabi3 Mario Fritz3
*Equal contribution
1University of Maryland 2Max Planck Institute for Informatics 3CISPA Helmholtz Center for Information Security
https://github.com/ningyu1991/ArtificialGANFingerprints

Motivations
• Generative models, especially Generative Adversarial Networks (GANs), evolve fast in the past 7 years for photorealistic generation, which raise significant concerns about visual misinformation.
• Move from passive to proactive defense against DeepFake misuses.
• Enable responsible release and regulation of DeepFake models.
• Be sustainable and independent of the arms race between DeepFake generation and detection.

Pipeline
• Incorporate unique artificial fingerprints into generators.
• Learn a fingerprint detector to detect fingerprints from generated images.
• Instantiate different generator versions with different fingerprints, each version deployed for a user download.
• Assume fingerprints transfer from training images, through generator, to generated images.

Training
Random Fingerprints $< c > (0,1)^N$

Deployment
User-specific Fingerprints

Detection
Detected Fingerprint

Fingerprint transferability and generation fidelity

Goals
Fingerprint transferability
Generation fidelity
Fingerprint robustness
Deepfake Detection/Attribution

DeepFake Detection and Attribution

Fingerprint Robustness
Gaussian noise Gaussian blur JPEG compression

Fingerprints encoded in the training images should present in the generated images
Encoded fingerprints do not hurt the original generation quality and keep invisible to human eyes
Encoded fingerprints should persist within a reasonably wide range of model variability
The effectiveness to convert the complex classification problem to the simple fingerprint verification problem

Method
• Train a fingerprint auto-encoder on the real training images.
• Apply the well-trained encoder to the entire training dataset.
• Train any generative model in the original way using the fingerprinted dataset.
• Detect fingerprints on the generated images from fingerprinted generators and match them to the fingerprint database.

Fingerprints transferability and generation fidelity

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Fingerprint tech</th>
<th>Model</th>
<th>Bit acc †</th>
<th>$p$-value</th>
<th>Orig FID</th>
<th>FID ‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>CelebA</td>
<td>Eq. 6</td>
<td>ProGAN</td>
<td>0.93</td>
<td>$&lt; 10^{-5}$</td>
<td>14.09</td>
<td>60.28</td>
</tr>
<tr>
<td></td>
<td>Eq. 6 [1]</td>
<td>StyxGAN2</td>
<td>0.51</td>
<td>$&lt; 10^{-5}$</td>
<td>6.41</td>
<td>6.91</td>
</tr>
<tr>
<td></td>
<td>Eq. 6 [1]</td>
<td>StyxGAN2</td>
<td>0.53</td>
<td>0.46</td>
<td>6.41</td>
<td>6.82</td>
</tr>
<tr>
<td></td>
<td>Eq. 6 [4]</td>
<td>StyxGAN2</td>
<td>0.98</td>
<td>$&lt; 10^{-5}$</td>
<td>8.98</td>
<td>9.72</td>
</tr>
<tr>
<td></td>
<td>Eq. 6 [4]</td>
<td>StyxGAN2</td>
<td>0.99</td>
<td>$&lt; 10^{-5}$</td>
<td>6.41</td>
<td>6.23</td>
</tr>
<tr>
<td>LSUN Around</td>
<td>Eq. 6</td>
<td>ProGAN</td>
<td>0.93</td>
<td>$&lt; 10^{-5}$</td>
<td>29.16</td>
<td>32.58</td>
</tr>
<tr>
<td></td>
<td>Eq. 6 [1]</td>
<td>StyxGAN2</td>
<td>0.98</td>
<td>$&lt; 10^{-5}$</td>
<td>24.95</td>
<td>25.71</td>
</tr>
<tr>
<td></td>
<td>Eq. 6 [1]</td>
<td>StyxGAN2</td>
<td>0.99</td>
<td>$&lt; 10^{-5}$</td>
<td>13.92</td>
<td>14.71</td>
</tr>
<tr>
<td>LSUN</td>
<td>Eq. 6</td>
<td>ProGAN</td>
<td>0.98</td>
<td>$&lt; 10^{-5}$</td>
<td>45.22</td>
<td>48.97</td>
</tr>
<tr>
<td></td>
<td>Eq. 6 [1]</td>
<td>StyxGAN2</td>
<td>0.99</td>
<td>$&lt; 10^{-5}$</td>
<td>33.45</td>
<td>34.01</td>
</tr>
<tr>
<td></td>
<td>Eq. 6 [1]</td>
<td>StyxGAN2</td>
<td>0.99</td>
<td>$&lt; 10^{-5}$</td>
<td>31.01</td>
<td>32.60</td>
</tr>
<tr>
<td>CIFAR-10</td>
<td>Eq. 6</td>
<td>BigGAN</td>
<td>0.99</td>
<td>$&lt; 10^{-5}$</td>
<td>6.25</td>
<td>6.80</td>
</tr>
<tr>
<td>Horse→Zebra</td>
<td>Eq. 6</td>
<td>CUT</td>
<td>0.99</td>
<td>$&lt; 10^{-5}$</td>
<td>22.98</td>
<td>23.43</td>
</tr>
<tr>
<td>Cat→Dog</td>
<td>Eq. 6</td>
<td>CUT</td>
<td>0.99</td>
<td>$&lt; 10^{-5}$</td>
<td>55.78</td>
<td>56.09</td>
</tr>
</tbody>
</table>

DeepFake generation

Real generation

Fingerprinted generation