Artificial Fingerprinting for Generative Models: Rooting Deepfake Attribution in Training Data

Ning Yu*[1,2] Vladislav Skripniuk*[3] Sahar Abdelnabi[3] Mario Fritz[3]

*Equal contribution

[1]University of Maryland [2]Max Planck Institute for Informatics [3]CISPA Helmholtz Center for Information Security

https://github.com/ningyu1991/ArtificialGANFingerprints

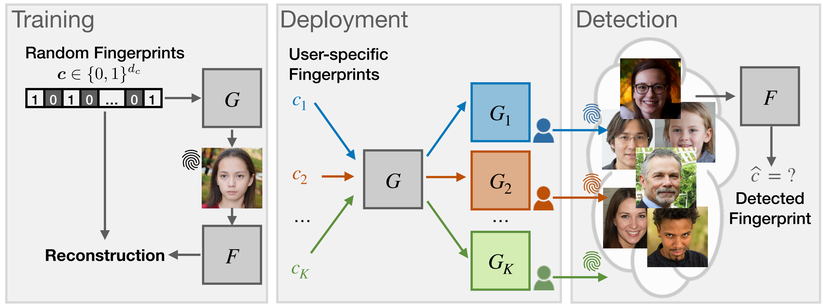
2021 ICCV OCTOBER 11-17 VIRTUAL

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** in 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.

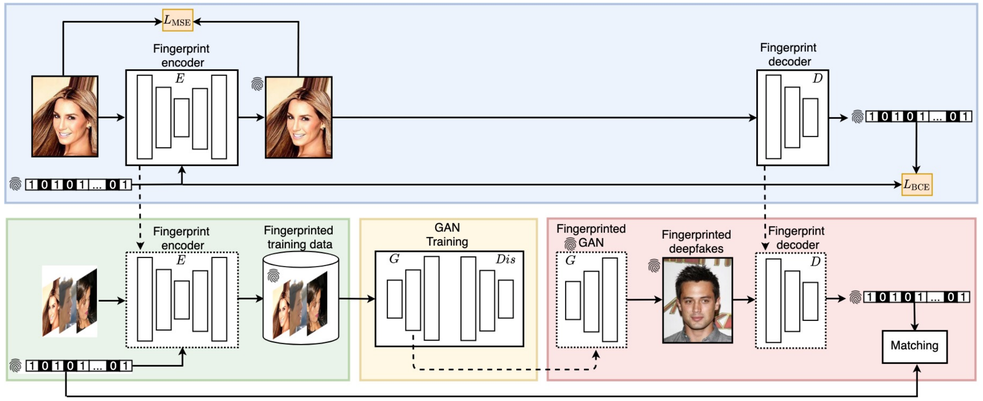


Goals

Fingerprint transferability	Generation fidelity	Fingerprint robustness	DeepFake Detection/attribution
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 image/model **perturbations**	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.



Fingerprint transferability and generation fidelity

Dataset	Fgpt tech	Model	Bit acc ⇑	p-value	Orig FID	Fgpt FID ⇓
CelebA	Eq. 6	ProGAN	0.93	$< 10^{-19}$	14.09	60.28
	[2]	StyleGAN2	0.51	0.46	6.41	6.93
	[1]	StyleGAN2	0.53	0.31	6.41	6.82
	[44]	Data	1.00	-	-	1.15
	[44]	ProGAN	0.98	$< 10^{-26}$	14.09	14.38
	[44]	StyleGAN	0.99	$< 10^{-28}$	8.98	9.72
	[44]	StyleGAN2	0.99	$< 10^{-28}$	6.41	6.23
LSUN *Bedroom*	[44]	ProGAN	0.93	$< 10^{-19}$	29.16	32.58
	[44]	StyleGAN	0.98	$< 10^{-26}$	24.95	25.71
	[44]	StyleGAN2	0.99	$< 10^{-28}$	13.92	14.71
LSUN *Cat*	[44]	ProGAN	0.98	$< 10^{-26}$	45.22	48.97
	[44]	StyleGAN	0.99	$< 10^{-28}$	33.45	34.01
	[44]	StyleGAN2	0.99	$< 10^{-28}$	31.01	32.60
CIFAR-10	[44]	BigGAN	0.99	$< 10^{-28}$	6.25	6.80
Horse→Zebra	[44]	CUT	0.99	$< 10^{-28}$	22.98	23.43
Cat→Dog	[44]	CUT	0.99	$< 10^{-28}$	55.78	56.09

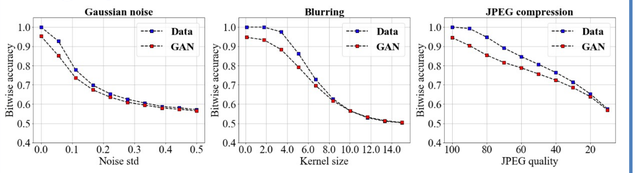
Original real 128x128

Fingerprinted real

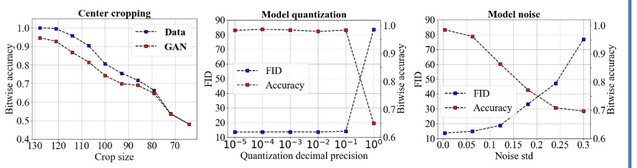
Original generation

Fingerprinted generation

Fingerprint Robustness



Gaussian noise | Gaussian blur | JPEG compression



Center crop | Model quantization | Model noise

DeepFake Detection and Attribution

Dataset	Model	Detector	Detection acc ⇑	Attribution acc ⇑
CelebA	ProGAN	[52]	0.508	0.235
		[49]	0.924	N/A
		Ours	1.000	1.000
	StyleGAN	[52]	0.497	0.168
		[49]	0.906	N/A
		Ours	1.000	1.000
	StyleGAN2	[52]	0.500	0.267
		[49]	0.895	N/A
		Ours	1.000	1.000
LSUN *Bedroom*	ProGAN	[52]	0.493	0.597
		[49]	0.952	N/A
		Ours	1.000	1.000
	StyleGAN	[52]	0.499	0.366
		[49]	0.956	N/A
		Ours	1.000	1.000
	StyleGAN2	[52]	0.491	0.267
		[49]	0.930	N/A
		Ours	1.000	1.000
LSUN *Cat*	ProGAN	[49]	0.951	N/A
		Ours	1.000	1.000
	StyleGAN	[49]	0.923	N/A
		Ours	1.000	1.000
	StyleGAN2	[49]	0.905	N/A
		Ours	1.000	1.000
CIFAR-10	BigGAN	[49]	0.815	N/A
		Ours	1.000	1.000
Horse→Zebra	CUT	[49]	0.836	N/A
		Ours	1.000	1.000
Cat→Dog	CUT	[49]	0.902	N/A
		Ours	1.000	1.000