Smudged Fingerprints

A systematic evaluation of the adversarial robustness of AI image fingerprints

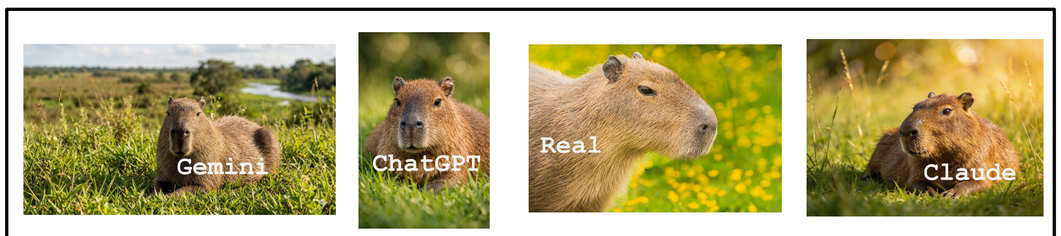
Kai Yao and Marc Juarez
University of Edinburgh | Poster @ IEEE SaTML 2026
Technical University of Munich, Germany

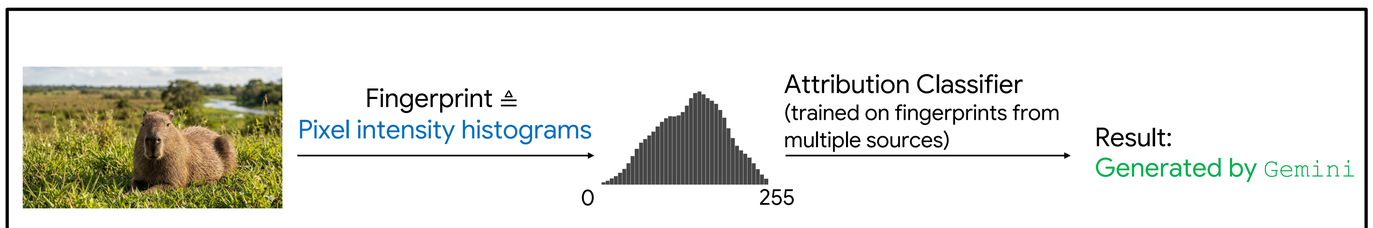
THE UNIVERSITY of EDINBURGH
THE UNIVERSITY of EDINBURGH
Generative AI Laboratory
IEEE

Model fingerprinting has been proposed for model attribution task; however, its adversarial robustness remains under-evaluated.

14	12	2	3 access levels -> 5 attacks	FFHQ 256x256
fingerprinting methods	image generators	attack goals	white-box to black-box	shared benchmark setting

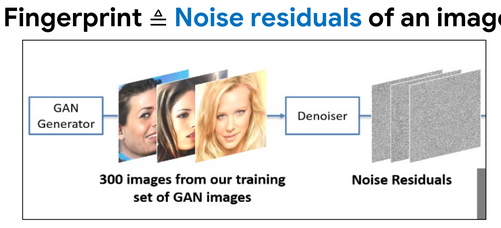
Model attribution (identifying generative AI sources) with fingerprinting





Model fingerprint examples

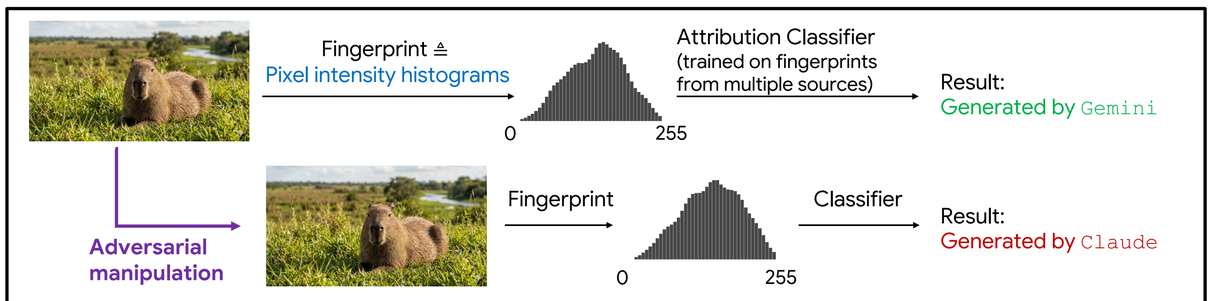
Fingerprint ≙ Noise residuals of an image



Fingerprint ≙ Pixel intensity histograms of an image

We study 14 fingerprinting methods that span 3 domains of image characteristics: **(1) RGB stats**: color and pixel dependencies; **(2) Frequency stats**: spectra, DCTs, residual structure; **(3) NN-Learned**: implicit high-dimensional fingerprints.

Clean attribution accuracy is not the full story

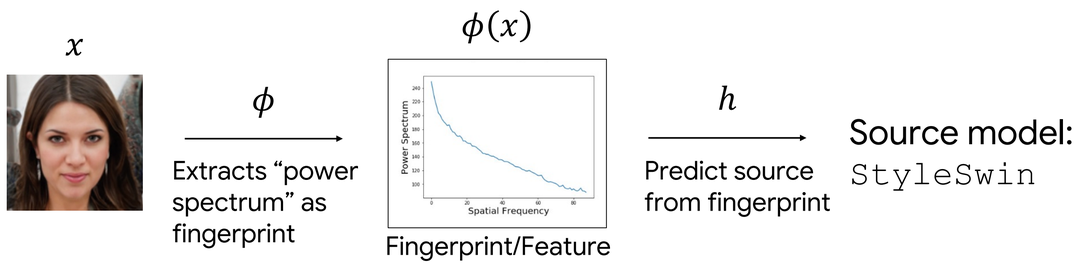


Our contributions: We formalize system and threat models, develop adaptive attacks in model attribution context, and conduct robustness benchmark on a large scale.

Model attribution system formulation

$$\mathcal{F}(x) = h\big(\phi(x)\big)$$

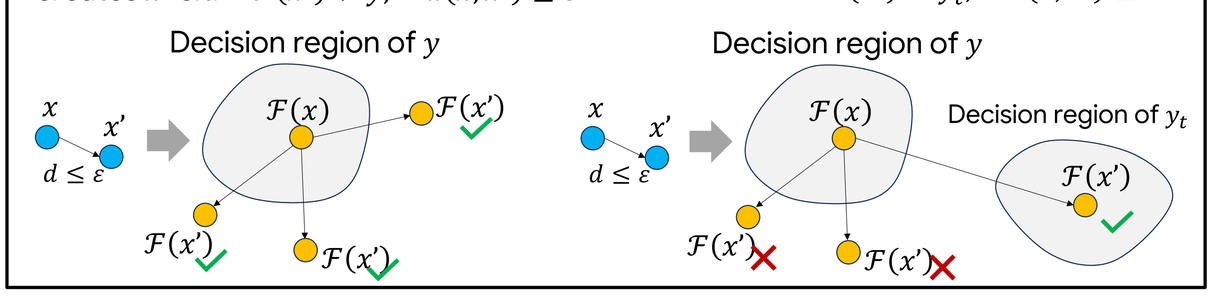
ϕ is the **fingerprint extractor**
h is the **model attribution classifier**



Attack goals

Removal: For $\mathcal{F}(x) = y$, the attacker creates x' s.t. $\mathcal{F}(x') \neq y, \quad d(x, x') \leq \varepsilon$

Forgery: For $\mathcal{F}(x) = y, y_t \neq y$, the attacker creates x' s.t. $\mathcal{F}(x') = y_t, \quad d(x, x') \leq \varepsilon$



Results: Removal is easy, even under black-box access

Fingerprint Method	Clean Acc	ASR (Attack Success Rate)		
		Best of W1~W3	B1	Best B2
— RGB				
McCloskey18	23.8%	59.4%	13.5%	13.3%
Nataraj19	91.0%	97.5%	84.8%	51.4%
Nowroozi22	87.3%	84.9%	84.5%	42.4%
Song24-RGB	63.8%	91.4%	2.3%	1.2%
— FREQUENCY				
Marra19a	66.7%	100.0%	9.6%	13.5%
Durall20	80.9%	99.9%	99.5%	94.1%
Dzanic20	79.0%	99.1%	99.8%	98.6%
Giudice21	94.9%	100.0%	92.7%	81.2%
Corvi23-R	39.6%	100.0%	35.2%	25.2%
Corvi23-S	27.1%	77.4%	15.1%	32.2%
Song24-Freq	84.2%	31.1%	70.6%	68.6%
— NN-LEARNED				
Qian20	90.1%	90.7%	82.0%	75.9%
Wang20	98.5%	100.0%	78.2%	49.0%
Song24-SL	87.9%	84.5%	55.9%	34.6%

Main results

- Due to lack of security designs, most methods exceed **80%** ASR in **white-box**.
- Even generic **B2** transforms (noising, blurring, JPEG, resizing) reach **50%** for many.

Practical concern

B1 remains competitive for several methods, so hiding detector is not enough – often knowing the attribution model pool is enough.

Most robust method

Song24-RGB (RGB manifold deviation) **Marra19a** (Denoising residuals)

Results: Forgery is harder, but still non-negligible

Fingerprint Method	Clean Acc	ASR (Attack Success Rate)	
		Best of W1~W3	B1
— RGB			
McCloskey18	23.8%	10.4%	2.1%
Nataraj19	91.0%	51.6%	11.3%
Nowroozi22	87.3%	12.1%	10.2%
Song24-RGB	63.8%	16.4%	0.0%
— FREQUENCY			
Marra19a	66.7%	97.8%	0.5%
Durall20	80.9%	57.7%	18.2%
Dzanic20	79.0%	25.5%	25.5%
Giudice21	94.9%	98.9%	13.9%
Corvi23-R	39.6%	99.8%	3.7%
Corvi23-S	27.1%	9.1%	2.5%
Song24-Freq	84.2%	3.3%	6.7%
— NN-LEARNED			
Qian20	90.1%	18.1%	12.5%
Wang20	98.5%	99.5%	5.7%
Song24-SL	87.9%	10.1%	7.3%

Why is forgery harder?

Removal only needs any wrong label. Forgery must land on one chosen target region in a 12-way attribution task.

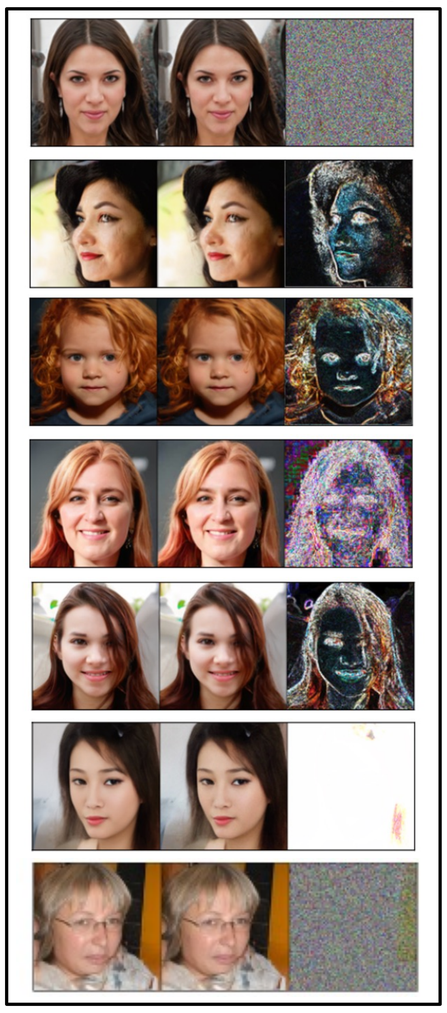
But it is still feasible

Several **white-box** attacks still approach **100%** forgery ASR. Black-box I reaches up to 25.51% (Dzanic20).

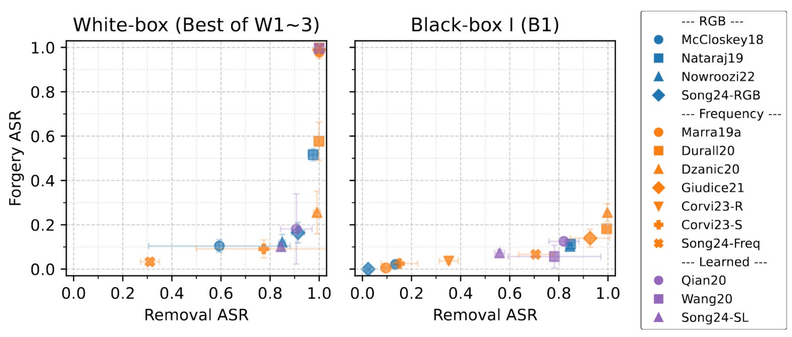
Most robust method

Song24-RGB (RGB manifold deviation) **Marra19a** (Denoising residuals) **Wang20** (ResNet50 features)

Example attack sample



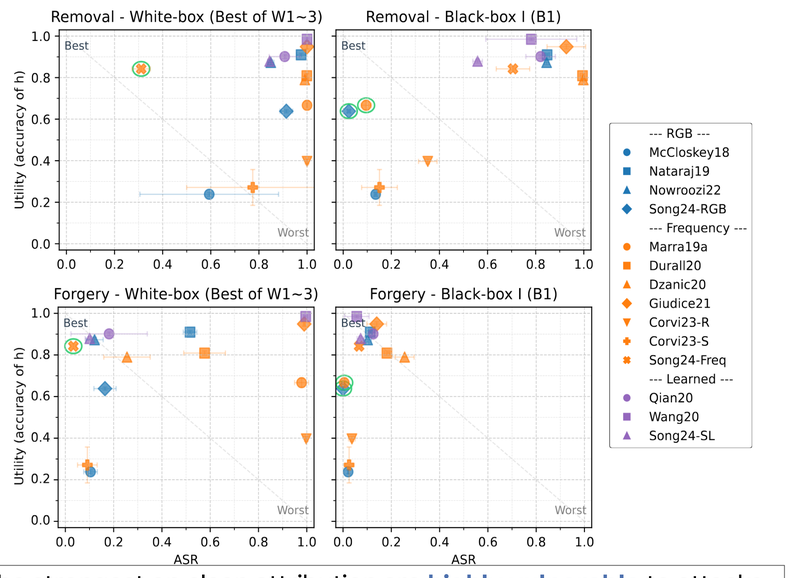
Removal and forgery ASR correlates



Removal-forgery correlation is clear

- If gradients can steer samples into a chosen forgery region, they can usually also push them out of the correct region.
- As explained, forgery is consistently harder than removal due to the size of model pool, hence the magnitude difference.

Utility—robustness tradeoffs are visible

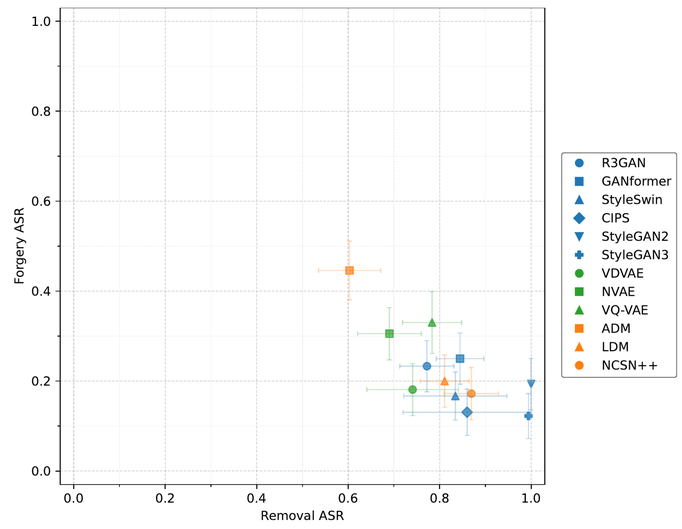


The strongest on clean attribution are **highly vulnerable** to attacks

Most promising black-box
Song24-RGB (RGB manifold deviation)
Marra19a (Denoising residuals)

Most promising white-box
Song24-Freq (Freq manifold deviation)

Attack variance across models



Models more resistant to removal tend to be more vulnerable to forgery, and vice versa.

Diffusion and VAE models trend more removal-resistant and more forgery-vulnerable; GANs trend the other way.