Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints

Ning Yu[1,2] Larry Davis[1] Mario Fritz[3]

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

https://github.com/ningyu1991/GANFingerprints

max planck institut informatik

CISPA HELMHOLTZ-ZENTRUM i. G.

ICCV 2019 Seoul, Korea

Motivations

- GAN challenges to **visual forensics** due to its increasingly appealing quality.
- GAN challenges to **intellectual property protection** due to the difficult task of attributing generated images to their GAN sources.

Problem Statement

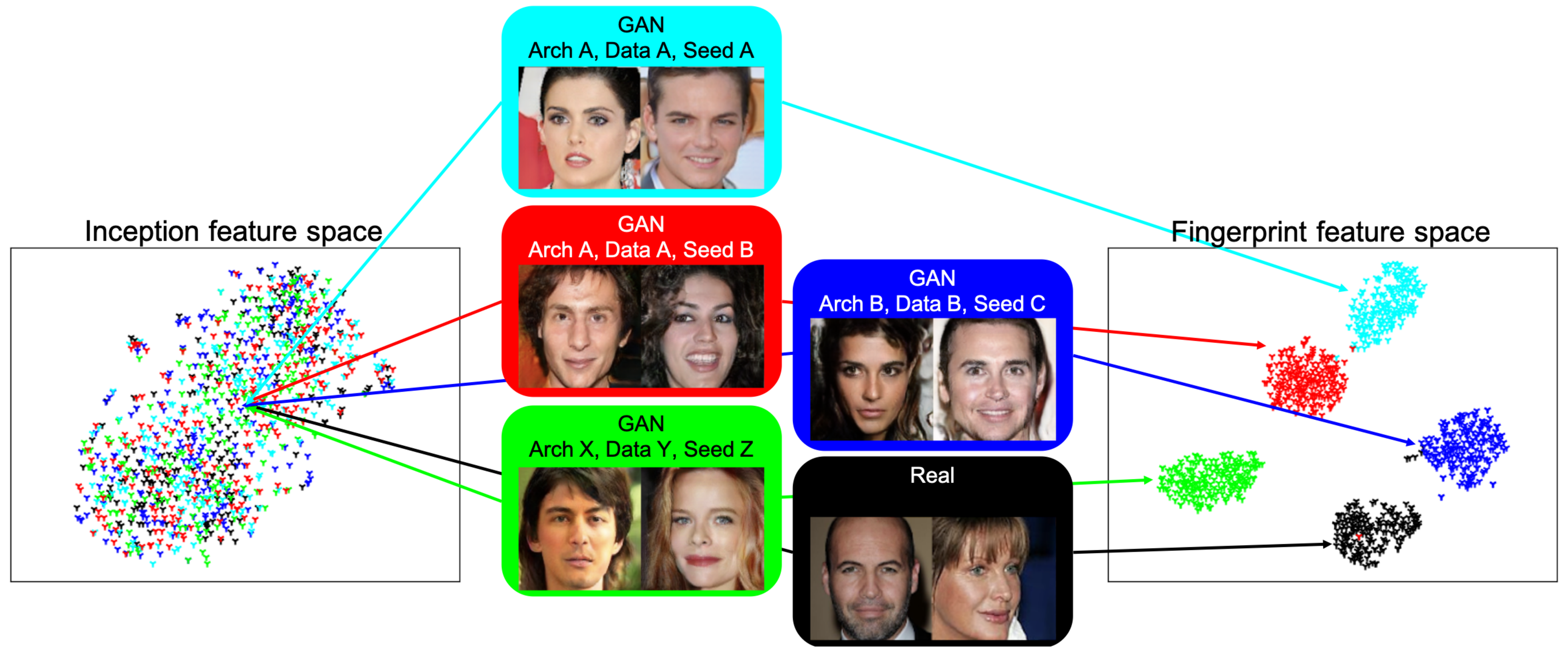
We address the two GAN challenges simultaneously by learning GAN fingerprints for image attribution: We introduce GAN fingerprints and use them to classify an image as real or GAN-generated. For GAN-generated images, we further identify their sources.

Fingerprints

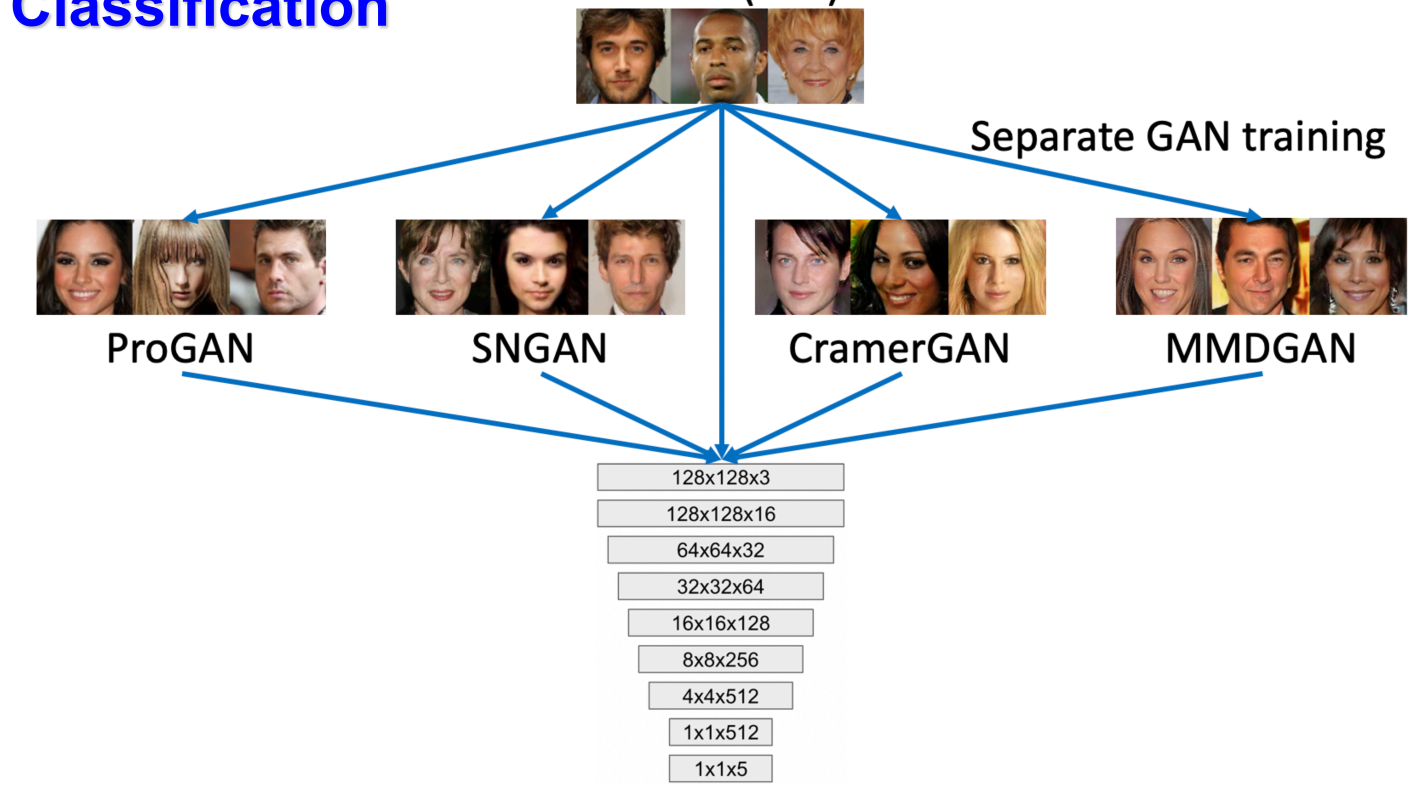
- **Model fingerprints**: We define the model fingerprint per GAN instance as a reference vector, such that it consistently interacts with all its generated images. E.g., parameters of the final fully-connected layer in an attribution classifier network.
- **Image fingerprints**: We define the fingerprint per image as a feature vector encoded from that image. E.g., features ahead of the final fully-connected layer in an attribution classifier network.

Insights

- **Existence**: GANs carry distinct model fingerprints and leave stable fingerprints in their generated images, which support image attribution.
- **Uniqueness**: Even minor differences in GAN training can result in different fingerprints, which enables fine-grained model authentication.
- **Persistence**: Fingerprints persist across different image frequencies and patches and are not biased by GAN artifacts
- **Immunizability**: Fingerprint finetuning is effective in defending against five types of image perturbation attacks.
- **Visualization**: We propose an alternative classifier variant to explicitly visualize GAN fingerprints in the image domain, so as to better interpret the effectiveness of attribution.
- **Superiority**: Comparisons also show our learned fingerprints consistently outperform several baselines in a variety of setups.



Classification

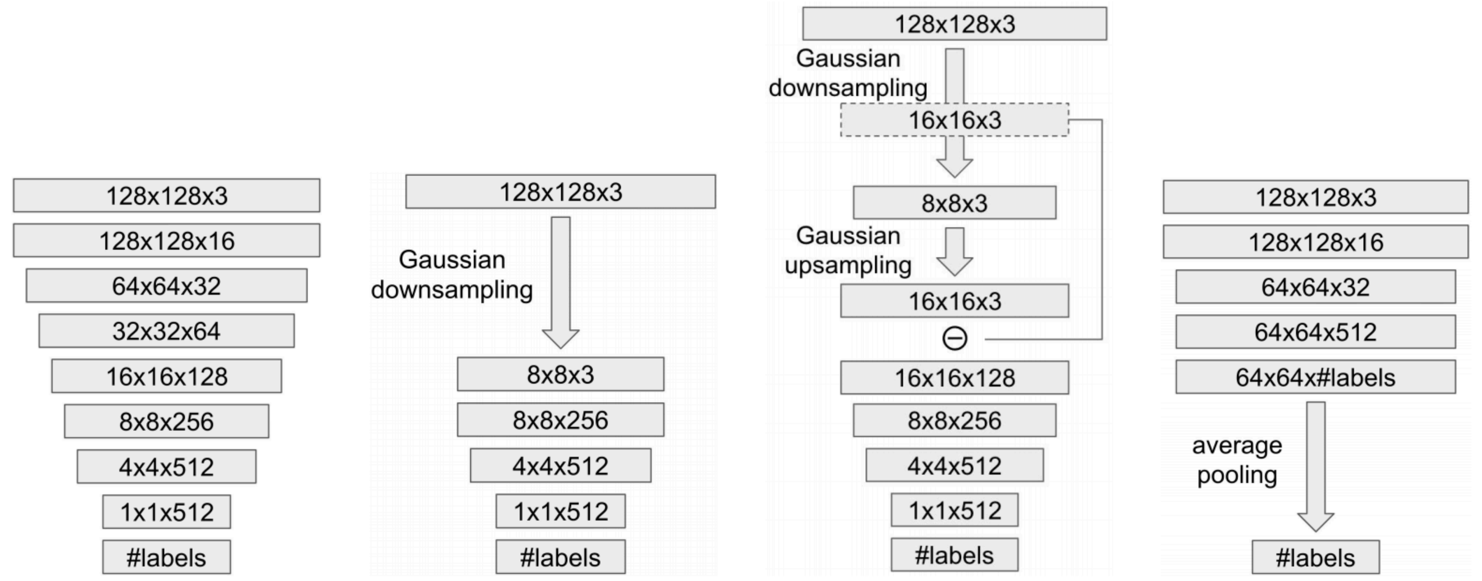


{CelebA (real), ProGAN, SNGAN, CramerGAN, MMDGAN}

		CelebA	LSUN
Accuracy (%)	kNN	28.00	36.30
	Eigenface	53.28	-
	PRNU	86.61	67.84
	Ours	**99.43**	**98.58**
	Our visNet	97.07	96.58
FD ratio	Inception	2.36	5.27
	Our fingerprint	**454.76**	**226.59**

$$\text{FD ratio} = \frac{\text{inter-class FD}}{\text{intra-class FD}}$$

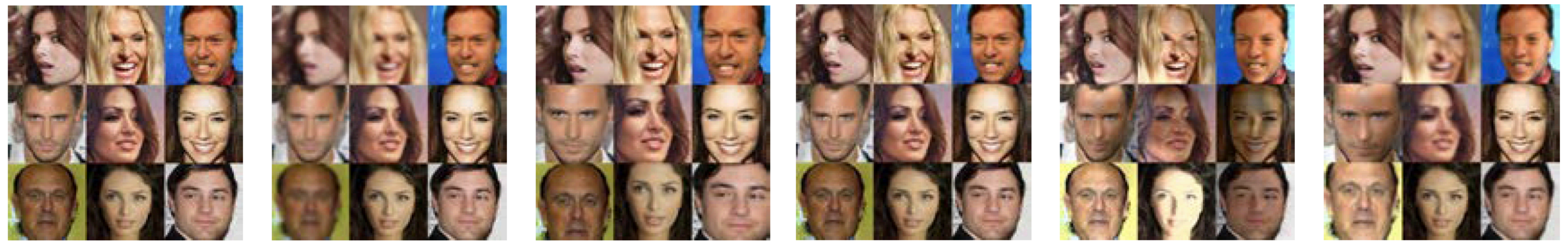
Classification Variants



Downsample factor	Resolution	CelebA L-f	CelebA H-f	LSUN L-f	LSUN H-f
1	128^2	99.14	99.14	97.04	97.04
2	64^2	98.74	98.64	96.78	96.84
4	32^2	95.50	98.52	91.08	96.04
8	16^2	87.20	92.90	83.02	91.58
16	8^2	67.44	78.74	63.80	80.58
32	4^2	26.58	48.42	28.24	54.50

Pooling starts at	Patch size	CelebA	LSUN
4^2	128^2	99.34	97.44
8^2	108^2	99.32	96.30
16^2	52^2	99.30	95.94
32^2	24^2	99.24	88.36
64^2	10^2	89.60	18.26
128^2	3^2	13.42	17.10

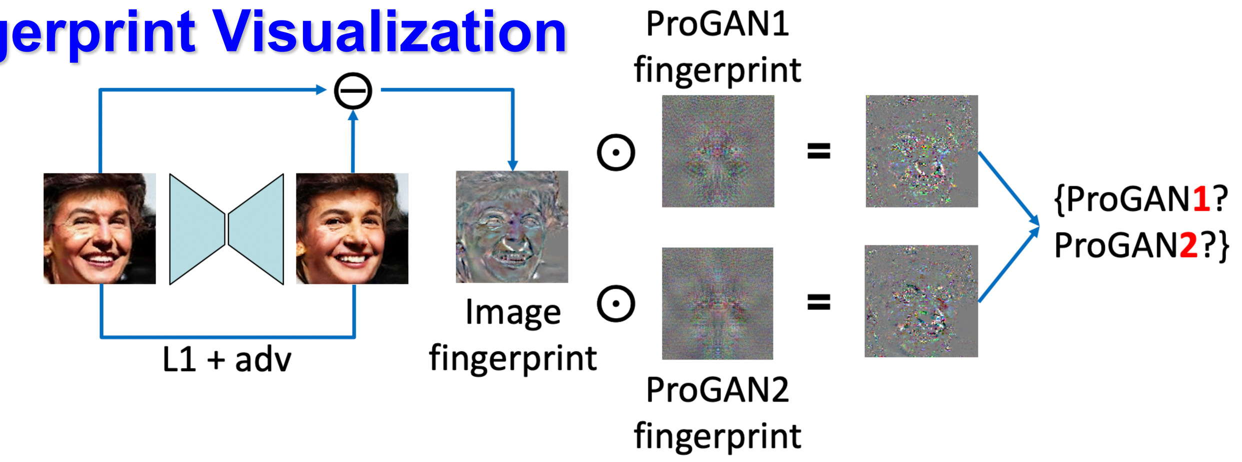
Attacks and Defenses



(b) Noise (c) Blur (d) Cropping (e) Compression (f) Relighting (g) Combination

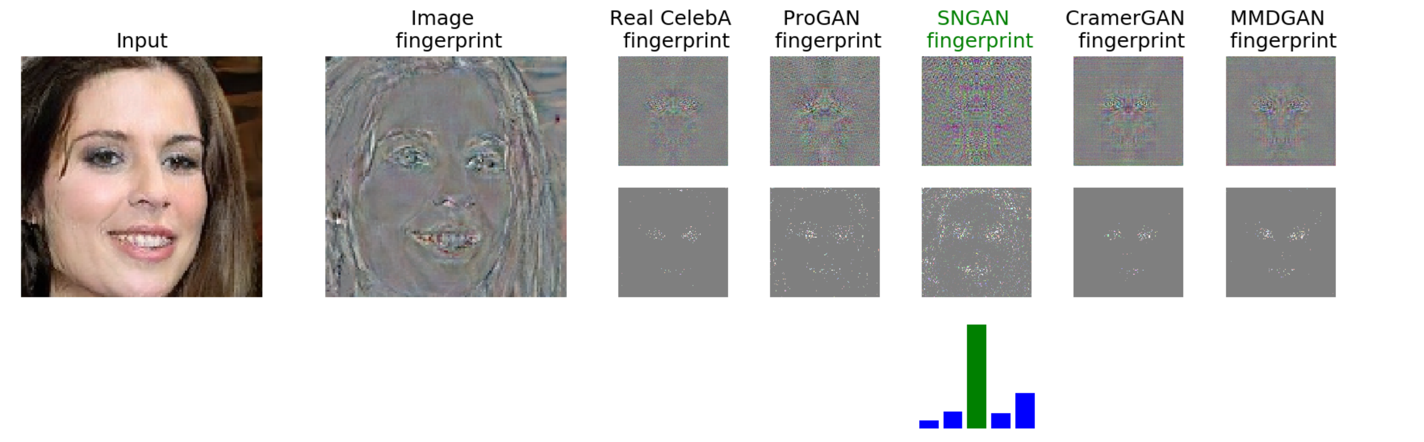
	Noise Atk	Noise Dfs	Blur Atk	Blur Dfs	Cropping Atk	Cropping Dfs	Compression Atk	Compression Dfs	Relighting Atk	Relighting Dfs	Combination Atk	Combination Dfs
	CelebA											
PRNU	**57.88**	63.82	27.37	42.43	9.84	10.68	**26.15**	44.55	86.59	87.02	**19.93**	21.77
Ours	9.14	**93.02**	**49.64**	**97.20**	**46.80**	**98.28**	8.77	**88.02**	**94.02**	**98.66**	19.31	**72.64**

Fingerprint Visualization

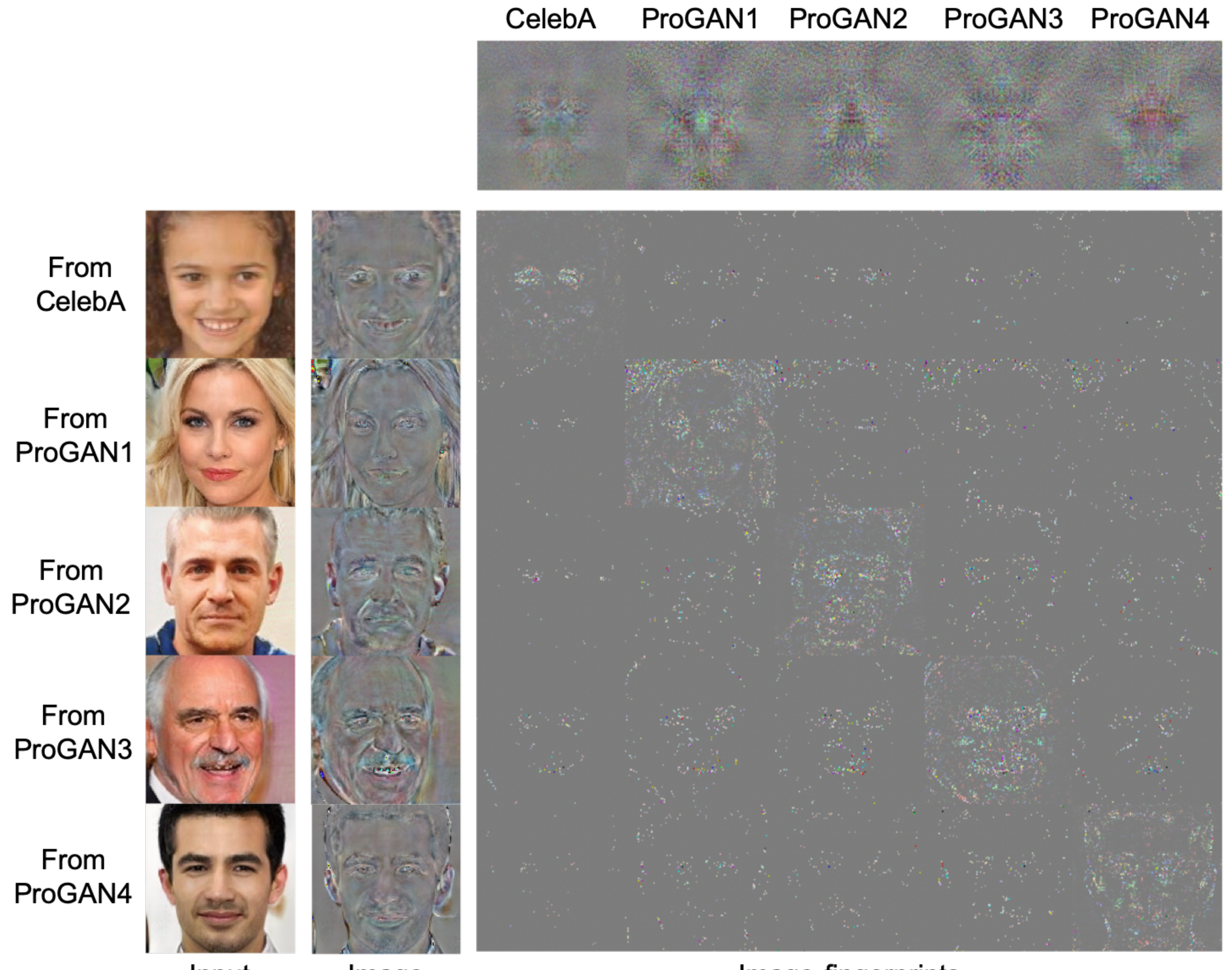


$$\min_{R,\{F^{\tilde y}_{mod}|\tilde y\in\mathbb{Y}\}} \max_{D_{rec}} \mathbb{E}_{\{(I,y)\}} (\lambda_1 L_{pix} + \lambda_2 L_{adv} + \lambda_3 L_{cls})$$

$$L_{cls}(I,y) = -\log \frac{corr(F^I_{im}, F^y_{mod})}{\sum_{\tilde y\in\mathbb{Y}} corr(F^I_{im}, F^{\tilde y}_{mod})}$$



Model fingerprints



Input images Image fingerprints Image fingerprints responding to model fingerprints